Texts in Statistical Science

Sampling

Design and Analysis

Third Edition

SOLUTIONS MANUAL

Sharon L. Lohr

Chapter 1

Introduction

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

Sampling frame: Persons who know about the online survey.
Sampling unit = observation unit: One response to the survey.

As noted in Section 1.3, samples that consist only of volunteers are suspect. This is especially true of surveys in which respondents must register and provide personal information to participate, as here. This survey is likely not intended to provide input to the editors, but rather to provide readers a way of feeling connected and involved with the magazine.

It also is possible that persons with strong opinions about the band BTS organized and contributed multiple responses to the poll.

1.2 Target population: All mutual funds.
Sampling frame: Mutual funds in the list.
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, or a book's author manages to flood the site with positive reviews.

1.4 Target population: Persons eligible for jury duty in Pennsylvania.
Sampling frame: State 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.

1.5 Target population: All persons experiencing homelessness 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 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: Likely voters in Iowa.
Sampling frame: Persons who attend the State Fair and stop by the booth.
Sampling unit = observation unit: One person.

This is a convenience sample 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.

1.9 Target population: Wikipedia science articles.
Sampling frame: Articles thought to be of interest by the sampler.
Sampling unit: One article.

This is a judgment sample, where the articles are deliberately chosen for review. The statistics apply only to the sample, and cannot be generalized to the population. There is additional possible selection bias because of the nonresponse; perhaps articles with many or few errors are harder to get reviewers for.

1.10 Target population: The 32,000 subscribers to the magazine.
Sampling frame: Magazine subscribers who saw the survey invitation.
Sampling unit: One response. Since more than 32,000 responses were received, it appears that either non-subscribers responded to the survey or some subscribers responded multiple times.

This is a convenience sample. The statistics have no validity because of the self-selected nature of the sample. A person who responded multiple times could unduly influence the survey results. In many surveys of this type, only persons with strong opinions respond. It's possible in this case that persons who have been inconvenienced by a PC malfunction will be more likely to respond than persons who have experienced no problems with their PCs.

1.11 Target population: Unclear. Presumably they want to generalize the results to adult women.
Sampling frame: Persons who see and respond to the online invitation.
Sampling unit: One response. It is possible that some persons responded multiple times.

This is a convenience sample. The statistics have no validity because of the self-selected nature of the sample. We don't even know if 4,000 different women responded, or even if the respondents are women; it could be just a handful of persons each responding multiple times.

1.12 Target population: Faculty members at higher education institutions.
Sampling frame: Faculty members in the purposively sampled institutions. The authors did not state how they selected the faculty in those institutions who were surveyed.
Sampling unit: One faculty member.

There are several sources of selection bias. The first is the purposive nature of the college sample. Although the authors stated that the institutions are "representative in terms of institutional type, geographic and demographic diversity, religious/nonreligious affiliation, and the public/private dimension," it is very easy for conscious or unconscious subjective biases to skew such a sample. A second source of potential bias comes from the selection of faculty to participate (if not every-

one or a random sample). Finally, only about 60% of the selected faculty members respond, so nonresponse is another potential source of bias.

1.13 Target population: All ASA members.
Sampling frame: ASA members with active e-mail addresses.
Sampling unit: One member.

ASA members could respond only if e-mailed the survey, and this ensured that each member responded at most once. But the low response rate raises concerns that the respondents may be persons who are strongly engaged in the topic.

1.14 Target population: All married couples in the U.S.
Sampled population: Married couples who were in therapy or counseling
Sampling unit: One counselor
Observation unit: One couple

Couples in therapy may be experiencing more marital problems than other couples.

1.15 Target population: All adults
Sampling frame: Friends and relatives of American Cancer Society volunteers
Sampling unit: One person

This is a convenience sample, and the results cannot be applied to the general population. 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." (Lohr, 2008, pp. 97–98)

1.16 Target population: Readers of romance fiction
Sampled population: Women who self-identified as readers of romance fiction
Sampling unit: One woman
Observation unit: One woman

This was almost certainly a convenience sample.

1.17 Target population: All statisticians

Sampled population: Statisticians whose names are in the online directory of a statistical organization (and whose e-mail address is correct)
Sampling unit: One person
Observation unit: One person

There are several potential sources of bias. First, not all statisticians belong to one of the organizations and often the e-mail addresses in those listings are out of date. Then, there is nonresponse bias. There is also a potential problem that a statistician belonging to more than one organization may be sampled multiple times.

1.18 Target population: All adults U.S.
Sampled population: Adults with telephone access
Sampling unit: Telephone number
Observation unit: Adult

There may be some undercoverage here, since not everyone has a telephone, but that is quite low. The main source of potential bias here is nonresponse. Fewer than 30% of the eligible numbers resulted in an interview. One can come up with potential reasons for the bias being in either direction.

1.19 Target population: Members of Consumers Union
Sampled population: Members with an e-mail address who would be willing to respond to a survey
Sampling unit: One member (could be household)
Observation unit: One member

The big problem here is nonresponse. It is reasonable to assume that persons with strong opinions would be more likely to respond to the survey.

1.20 Target population: All students in the state
Sampled population: School districts in the state with internet access
Sampling unit: One school
Observation unit: One student

This is a purposive sample of school districts, and a convenience sample of students within schools.

1.21 Target population: All brains of deceased football players
Sampled population: Brains that had been donated to the brain bank
Sampling unit: One brain
Observation unit: One brain

This is not a representative sample because family members were probably more likely to donate their loved ones' brains for scientific research if they had shown symptoms of post-concussive injury. You cannot conclude from this study that 99% of NFL players will get CTE, even though many

newspaper articles made that implication.

The authors of the study, in Mez et al. (2017), were careful not to draw any implications about the population. One of their purposes was to correlate CTE with the amount of brain injury during life, and they found that men who had played football longer had, on average, worse brain damage. Again, this finding cannot necessarily be generalized to the population of all football players because it's possible that the relationship between time spent playing football and brain damage is stronger (or weaker) in this non-representative sample than in the population. But the results from the study are illuminating, and suggest that repeated brain trauma should be avoided.

1.22 Target population: Not clear. All people? Readers?
Sampled population: Persons willing to e-mail the columnist
Sampling unit: One reader
Observation unit: One reader

This is a pure convenience sample, and worthless for generalizing to any population. And what was to stop a reader from sending multiple responses?

1.23 Target population: All patients of the clinic
Sampled population: All patients of the clinic
Sampling unit: One patient
Observation unit: One patient

This likely gives a representative sample because patients are sampled as they enter. If patients arrive in random order, each has the same probability of being sampled. There might be some problems, if, say, patients arrive in busses that hold exactly 20 patients. Then it's possible that the first person off every bus would be chosen, who might differ from others.

1.24 Target population: All patients of the clinic
Sampled population: All patients of the clinic
Sampling unit: One patient
Observation unit: One patient

This procedure might give a biased sample because the next patient to leave is more likely to have had a longer appointment. Several patients with short appointments may leave while the interviewer is conducting the interview.

Geldsetzer et al. (2018) discuss different options for taking a sample of patients for exit interviews.

1.25 Target population: All watchers of the show
Sampled population: Persons who want to express an opinion about the contestants. As long as they have the number to text or the website address, they do not need to have watched
Sampling unit: One text or online submission

Observation unit: One text or online submission

This is a sample of convenience. According to `dwts.vote.abc.go.com/#faq`, accessed in January 2020, participants were allowed to submit up to 10 votes per couple online, and an additional 10 votes by text. The online votes had to be submitted through an ABC television account. But people with multiple cell phones could still submit multiple responses, and an organization could have members flood the line with responses.

Before 2019, the couple with the lowest combined score, found by averaging the audience vote score with the average judges' score, was eliminated from the competition. Because of complaints that determined fans were keeping weak dancers in the competition, the procedure changed in 2019. The judges chose one of the two lowest-scoring couples to eliminate.

1.26 Target population: All watchers of the news broadcast
Sampled population: Persons who go to the website to answer
Sampling unit: One online submission
Observation unit: One online submission

Of course this survey is purely entertainment, and the television station is careful to say that the poll is unscientific—one cannot conclude from this poll that 10% of all Arizonans think the state should participate in daylight savings time. This is a convenience sample.

1.27 Target population: All county election websites in Texas
Sampled population: All county election websites in Texas
Sampling unit: One county website
Observation unit: One county website

This is a census of all websites. Assuming there were no measurement errors, these statistics are accurate, with no sampling or nonsampling error. If the goal had been solely to estimate the percentage of county election websites that began with "http" rather than "https," it might have been easier to take a sample. But with a goal of obtaining an estimate and also identifying counties that do not use https, a census is needed.

1.28 Target population: All adults in the world who do not have children
Sampled population: People who hear about the survey
Sampling unit: One person

This is a pure convenience sample, and worthless for measuring opinion about nonparenting adults. Those who visit the author's website are likely to be persons who are already in agreement with the author's views, and so will reinforce the author's opinion.

1.29 Target population: Every U.S. physician
Sampled population: U.S. physicians with e-mail addresses in the file

Sampling unit: One physician
Observation unit: One physician

The low response rate for this survey is the biggest source of bias.

1.30 Target population: Residents of Little Rock
Sampled population: Residents of Little Rock who learn about the survey
Sampling unit: One resident
Observation unit: One resident

This is a convenience sample, and it is impossible to say whether this group is representative of the vast majority of Little Rock residents who do not respond. For most surveys of this type, people with strong opinions are the ones who will be willing to take the time to fill out the survey.

1.32 Target population: Residents of Guelph
Sampled population: Residents of Guelph who visited one of the in-person sites or the web site
Sampling unit: One resident (possibly with multiple visits?)

Both of these are convenience samples. The in-person survey will miss those who do not visit a site, and the online survey will miss those who do not learn about the survey or who do not choose to participate.

1.32

(a) How many potatoes did you eat last year?

How many people keep track of the number of potatoes they eat (unless they never eat potatoes)? Also, what counts as eating a potato? Do French fries count? How many potatoes are in a serving of scalloped or mashed potatoes?

(b) According to the American Film Institute, *Citizen Kane* is the greatest American movie ever made. Have you seen this movie?

This is a leading question; after being told that it is the greatest American movie ever made, respondents will be loath to say they have not seen it.

(c) Do you favor or oppose changing environmental regulations so that while they still protect the public, they cost American businesses less and lower product costs?

This leading question is from Kinsley (1981).

(d) Agree or disagree: We should not oppose policies that reduce carbon emissions.

Confusing because of double negative.

(e) How many of your close friends use marijuana products?

This is vaguely worded. What counts as a close friend? And what counts as a marijuana product?

(f) Do you want a candidate for president who would make history based on race, gender, sexual orientation, or religion?

This question appeared in *The New York Times* on January 30, 2020, see `https://www.nytimes.com/interactive/2020/01/30/us/politics/democratic-presidential-candidates-quiz.html`. The question is not just double-barreled, it is quadruple-barreled. And it is vague: What does it mean to make history?

(g) Who was or is the worst president in the history of the United States?

People don't remember about James Buchanan, and tend to vote for the current or a recent president.

(h) Agree or disagree: Health care is not a right, it's a privilege.

Asks two questions, not one. Answers will differ if the options are put in reverse order. And the agree/disagree format encourages people to agree.

(i) The British Geological Survey has estimated that the UK has 1,300 trillion cubic feet of natural gas from shale. If just 10% of this could be recovered, it would be enough to meet the UK's demand for natural gas for nearly 50 years or to heat the UK's homes for over 100 years. From what you know, do you think the UK should produce natural gas from shale?

Leading question.

(j) What hours do you usually work each day?

Presupposes that the person works.

(k) Do you support or oppose gun control?

A question such as "Do you support or oppose gun control?" will tell you little about public opinion because the term "gun control" can mean different things to different people. Some may think the term means requiring background checks before purchase, while others may think gun control prohibits all gun ownership by private citizens. A statistic from such a vague question cannot be interpreted without assuming how respondents interpret the question.

(l) Looking back, do you regret having moved so rapidly to be divorced, and do you now feel that had you waited, the marriage might have been salvaged?

Leading question.

(m) Are the Patriots and the Packers good football teams?

Asks two questions, not one. Also, what does it mean to be a good team?

(n) Antarctic ice is now melting three times faster than it was before 2012. And the Arctic is warming at twice the rate of the rest of the world. On a scale from 1 to 5, with 5 being the highest, how alarmed are you by melting ice, the rise of global sea levels, and increased flooding of coastal cities and communities?

This was a question from an Environmental Defense Fund survey that was mailed to the author. Who will not be alarmed after reading the introduction to the question? Not surprisingly, the last question on the survey asks for a donation of money.

(o) Overall, how strong was your emotional reaction to the performance of *La Bohème*? (No emotional response, Weak response, Moderate response, Strong response, or Not applicable)

What is an emotional reaction? Also, if this question is intended to measure satisfaction, someone could have a "strong emotional response" because the soprano was consistently off-key or the orchestra was terrible.

(p) How often do you wish you hadn't gotten into a relationship with your current partner? (All the time, Most of the time, More often than not, Occasionally, Rarely, Never)

Presumes there is a current partner and is vague. This question is from Stasko (2015, p. 90).

(q) Do you believe that for every dollar of tax increase there should be $2.00 in spending cuts with the savings earmarked for deficit and debt reduction? Double-barreled and leading

(r) There probably won't be a woman president of the U.S. for a long time and that's probably just as well.

This question is cited by Frankovic (2008); it is double-barreled.

(s) How satisfied are you that your health insurance plan covers what you expect it to cover? (Very satisfied, Satisfied, Dissatisfied, Very dissatisfied, or Not applicable)

Should ask what you expect the plan to cover as a separate question. This question was from a State of Arizona retirement survey.

(t) Given the world situation, does the government protect too many documents by classifying them as secret and top secret?

Is the question about the dire world situation, or about the government classifying too many documents?

1.34 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.

1.35 (a) There are several problems with the Landers survey. First, the sample is self-selected—only those who care enough to write a letter (and in 1975, this meant typing or handwriting it, and then putting it in an envelope with a stamp) would respond. Another problem is the leading question, which follows the story of the couple who regrets having children.

Note that Landers continued to cite results from the survey long after it had been discredited. In a 1996 book, she recalled the survey as follows (note that the quotes from the letter are different here):

> On November 3, 1975, a young married couple wrote to say they were undecided as to whether to have a family. They asked me to solicit opinions from parents of young children as well as from older couples whose families were grown. "Was it worth it?" they wanted to know. "Were the rewards enough to make up for the grief?" The question, as I put it to my readers, was this: "If you had it to do over again, would you have children?"
>
> Well, dear friends, the responses were staggering. Much to my surprise, 70 percent of those who responded said no. (Landers, 1996, p. 107)

Landers ran many surveys, as they stirred up reader interest and participation. For example, Landers (1996) described a survey on "divorce regret" which she took in 1993:

> In 1993, I asked my readers this question: "Looking back, do you regret having moved so rapidly to be divorced, and do you now feel that had you waited, the marriage might have been salvaged?"
>
> I asked for a yes or no answer on a postcard, but thousands of readers felt compelled to write long letters. I'm glad they did. I learned a lot.
>
> To my surprise, out of nearly 30,000 responses, almost 23,000 came from women. Nearly three times as many readers said they were glad they divorced, and most of them said they wished they had done it sooner. (pp. 14-15)

For another example, on Valentine's Day 1977, Landers asked readers to send a postcard answering the question: "If you had to do it over again, would you marry the person to whom you are now married?" She received more than 50,000 pieces of mail. 30% signed their cards; of those, 70% said they would marry the same person again. 80% of the signed cards came from women. The other 70% did not sign their cards; of those, 48% said yes and 70% were female.

The problem with a self-selected survey such as the one taken by Landers (1976) is that it continues to be cited as reliable information when in fact the survey results represent only the persons who happened to respond to this survey. For example, Dvorak (2011) cited the survey in an article in the *Washington Post*. And many news organizations cite such surveys precisely because they have eye-catching results. Once an unreliable statistic is out there, it's difficult to retract it.

(b) Again, the GH survey is self-selected, but it may have elicited a different response from readers who wanted to overturn the previous survey. Also, the readership of GH may differ from that of newspapers.

(c) Main questions: Was the sample selected randomly? And what was the response rate?

Note: I obtained the Newsday survey from Bryan Caplan's blog at `https://www.econlib.org/archives/2009/02/parents_and_buy.html` on February 2, 2020.

Another poll, in 2013, asked adults aged 45 and over: "If you had it to do over again, how many children would you have, or would you have no children at all?" and reported that 11% would have no children (Newport and Wilke, 2013). This was based on a "random sample of 5100 adults" conducted by telephone, with both landline and cellular telephones included.

Chapter 2

Simple Probability Samples

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

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

(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 following 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 (2.12):

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

(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 ith 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 2.30),

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

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

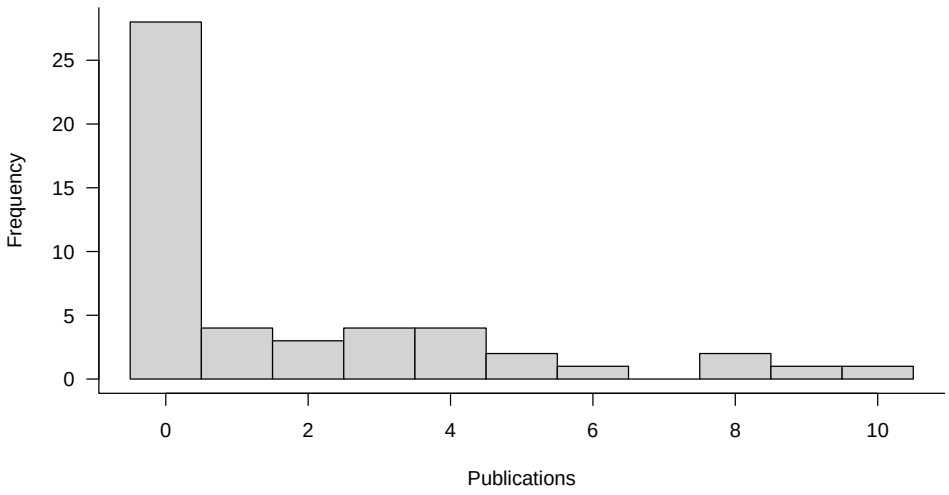
$$\text{SE}\,(\hat{t}) = \sqrt{3728.238} = 61.0593$$

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 software) 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 that this estimate differs from the estimate given by the UNIVARIATE procedure since the UNIVARIATE procedure adjusts for df using the formula

$$\text{skewness } = \frac{n}{(n-1)(n-2)} \sum_{i \in \mathcal{S}} (y_i - \bar{y})^3 / s^3.$$

Whichever estimate is used, however, formula (2.29) 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

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

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

$$\frac{175}{1000} \pm 1.96\sqrt{\frac{\frac{175}{1000}\left(1 - \frac{175}{1000}\right)}{1000}} = [0.151, 0.199].$$

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

2.8 (a) A stratified sample would be a better choice because the sampling frame has a lot of information about the population that could improve the efficiency of the design.

(b) A cluster sample would be a better choice. There is no list of patients, and they can only be accessed through their providers. First sample the allergists, then take a subsample of the patients of each sampled doctor.

(c) This one depends on how the website is organized. If there is an index where topics are organized A-Z, an SRS might be appropriate. Some websites are organized by links, however, or with subpages. In that case, a cluster sample may be better. Or this may be a network sample, as discussed in Chapter 14.

(d) For this sample, an SRS is likely the design of choice. The results of the audit need to be accepted by people on either side of the issue, and the only type of sample that many people think is fair is one that is selected completely randomly.

2.9 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 margin of error equals the t or normal critical value (approximately 2) times the standard error, $\text{SE}\,(\hat{p}) = \sqrt{p(1-p)/n}$. For a fixed sample size n, $p(1-p)$ is largest when $p = 1/2$. This can be shown through calculus since

$$\frac{d\,p(1-p)}{dp} = 1-2p$$

and $1-2p = 0$ when $p = 1/2$. Alternatively, a graph of $p(1-p)$ vs. p will show the maximum at $p = 1/2$.

(b) Table SM2.1 shows the margins of error for $p = 0.5$ in the first column. The second and third columns give the margin of error when $p = 0.2$ and $p = 0.1$, respectively. These values were calculated without the fpc.

(c) Values are given in Table SM2.2. The fpc makes almost no difference for most of the sample sizes except when the sample size is more than 10% of the population.

2.12 Assume the maximum value for the variance, with $p = 0.5$. Then use $n_0 = 1.96^2(0.5)^2/(.04)^2$, $n = n_0/(1+n_0/N)$. Table SM2.3 gives the results. The finite population correction makes a small difference for Tempe and Casa Grande, and a larger difference for Gila Bend and Jerome.

2.13 The sampling weight is 20,000/200 = 1,000. Answers will vary for the remaining parts.

Table 2.1: Margin of error (MOE) for an estimated proportion from an SRS, no fpc.

	MOE if Estimated Proportion is		
Sample Size	0.5	0.2 or 0.8	0.1 or 0.9
50	0.142098	0.113679	0.085259
100	0.099211	0.079369	0.059527
500	0.043933	0.035146	0.02636
1000	0.031027	0.024822	0.018616
5000	0.013862	0.01109	0.008317
10000	0.009801	0.007841	0.005881
50000	0.004383	0.003506	0.002630
100000	0.003099	0.002479	0.001859
1000000	0.00098	0.000784	0.000588

Table 2.2: Margin of error (MOE) for an estimated proportion from an SRS from a population with $N = 1$ million.

	MOE if Estimated Proportion is		
Sample Size	0.5	0.2 or 0.8	0.1 or 0.9
50	0.142095	0.113676	0.085257
100	0.099206	0.079365	0.059524
500	0.043922	0.035137	0.026353
1000	0.031012	0.024809	0.018607
5000	0.013828	0.011062	0.008297
10000	0.009752	0.007802	0.005851
50000	0.004272	0.003417	0.002563
100000	0.00294	0.002352	0.001764
1000000	0	0	0

Table 2.3: Results for Exercise 2.12.

City	Population	n_0	n
Casa Grande	48,571	600.25	593
Gila Bend	1,922	600.25	458
Jerome	444	600.25	256
Phoenix	1,445,632	600.25	600
Tempe	161,719	600.25	598

2.14 (a) The estimated proportion is $23/30 = 0.767$, or 76.7%. A 95% confidence interval for the

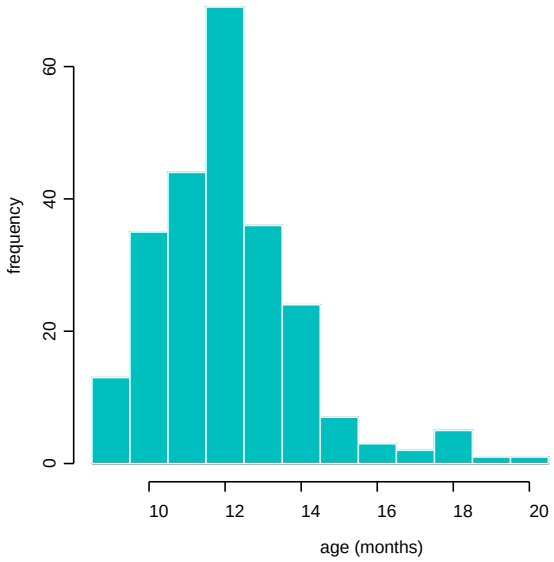
proportion of database members with a third cousin or closer relative in the database is

$$\frac{23}{30} \pm 2.045\sqrt{\frac{1}{30}\frac{23}{30}\left(1-\frac{23}{30}\right)} = [0.606, 0.927]$$

The confidence interval for the percentage is from 61% to 93%. Note that the fpc for this sample is negligible.

(b) The confidence interval in (a) applies *only* to the persons in the database, because that is where the sample was selected. They may well have a different matching percentage than persons in the public at large. Perhaps relatives tend to sign up for genetic testing together, which would increase the chances that a sampled person would have a close relative in the database.

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

2.16 (a) Using (2.30) 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% CI for the proportion of children overdue for vaccination is

$$\frac{93}{120} \pm 1.96\sqrt{\frac{\frac{27}{120}\left(1 - \frac{93}{120}\right)}{120}} = [0.70, 0.85].$$

2.17 (a) We have $\hat{p} = 0.2$ 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.

2.18 (a) There is no selection bias from taking the sample, since it is an SRS from the directory. There may be undercoverage, though, if the online directory does not include everyone.

(b) 30.7% female with 95% CI [23.9, 37.5]

(c) 265 female members with 95% CI [206.3, 323.6]

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

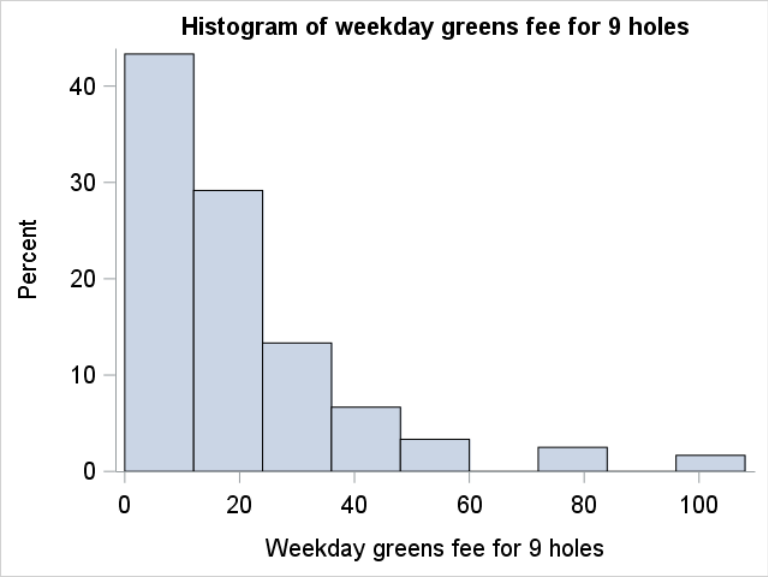
$$\text{CI}: 301953.7 \pm 1.96\sqrt{\frac{s^2}{300}\left(1 - \frac{300}{3078}\right)}, \text{ or } [264883, 339025]$$

(b) $\bar{y} = 599.06$, $s^2 = 161795.4$, CI : $[556, 642]$

(c) $\bar{y} = 56.593$, $s^2 = 5292.73$, CI : $[48.8, 64.4]$

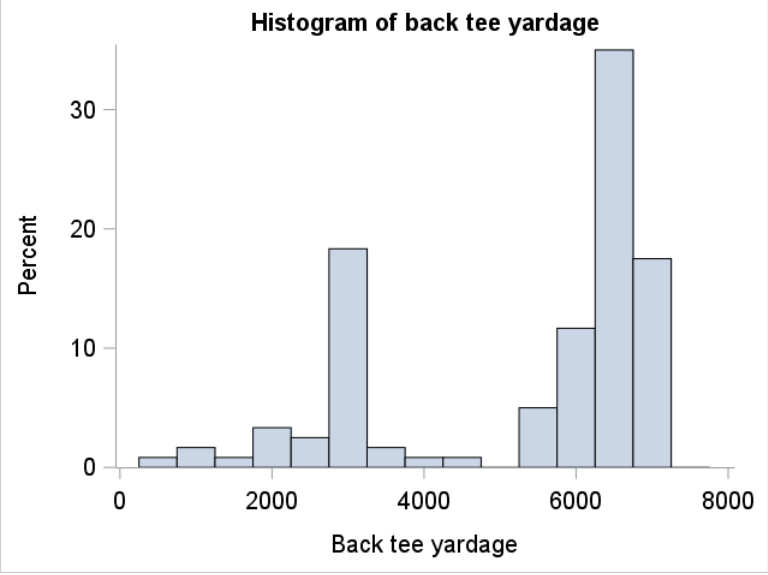
(d) $\bar{y} = 46.823$, $s^2 = 4398.199$, CI : $[39.7, 54.0]$

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



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

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



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

2.22 $\hat{p} = 85/120 = 0.708$

$$95\%\text{CI: } 85/120 \pm 1.96\sqrt{\frac{85/120\,(1-85/120)}{119}\left(1-\frac{120}{14938}\right)} = .708 \pm .081, \text{ or } [0.627, 0.790].$$

2.23 (a) The sampling weight is 7,048,107/5,000 = 1409.621.

(b) 27.66%, with 95% CI [26.42, 28.90]

(c) 415838 with 95% CI [369807.26, 461869.36]

(d) 13.1% with 95% CI [12.16, 14.034] The CV is 0.036415.

2.24 Sixty of the 70 samples yield confidence intervals, using this procedure, that include the true value $t = 40$. The exact confidence level is $60/70 = 0.857$, which is lower than 95%.

2.25 (a) From (2.16),

$$\text{CV}(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}\bar{y}_{\mathcal{U}}}.$$

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

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

The CV for a sample of size 1 is $\sqrt{(1-p)/p}$. The sample size in (2.32) will be $z^2_{\alpha/2}\text{CV}^2/r^2$.

(b) I used a spreadsheet to calculate these values.

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

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

2.26 *Decision theoretic approach for sample size estimation.*

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

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

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.

2.27 The supplementary books by Lohr (2022) and Lu and Lohr (2022) give code in SAS and R software, respectively, that may be used for this problem. The distribution from a sample of size 100 appears skewed.

2.28 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 kth unit is selected, the sample cannot consist of the first n units in the population.

2.29 (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 $x \log(0.99999) = \log(0.5)$, 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?"

2.30 (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):

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

and

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

(b)

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

(c)

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

2.31 (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 $\{1, \ldots, N\}$; the random variable $(U_{T_2} \mid T_1, \ldots, T_N, U_{T_1})$ 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

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

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,

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

$$
\begin{aligned}
&= \frac{1}{800}\sum_{l=0}^{\infty}\left[\sum_{k=1}^{20}\frac{1}{20}\frac{40-M_k}{40}\right]^l \\
&= \frac{1}{800}\sum_{l=0}^{\infty}\left[1-\frac{K}{800}\right]^l \\
&= \frac{1}{800}\frac{1}{1-(1-K/800)} = \frac{1}{K}.
\end{aligned}
$$

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.

2.32 Other proofs are possible (including by induction), but the following shows directly that the probability that a subset is chosen as the sample is that from an SRS.

We wish to show that for any subset of units $\mathcal{S} = \{a_1, a_2, \ldots, a_n\}$ from $\{1, \ldots, N\}$,

$$P(\mathcal{S}) = n!(N-n)!/N!.$$

Without loss of generality, let $a_1 < a_2 < \ldots < a_n$ and let $a_0 = 0$. Then

$$
\begin{aligned}
P(\mathcal{S}) = P\,&(\text{do not select units } a_0+1, \ldots, a_1-1, \text{ select unit } a_1, \\
&\quad \text{do not select units } a_1+1, \ldots, a_2-1, \text{ select unit } a_2, \ldots \\
&\quad\quad \text{do not select units } a_{n-1}+1, \ldots, a_n-1, \text{ select unit } a_n) \\
&= \prod_{j=0}^{n-1}\left\{\left[\prod_{k=a_j+1}^{a_{j+1}-1}\left(1-\frac{n-j}{N-k+1}\right)\right]\left[\frac{n-j}{N-a_{j+1}+1}\right]\right\} \\
&= \frac{\prod_{j=0}^{n-1}\left\{\left[\prod_{k=a_j+1}^{a_{j+1}-1}(N-k+1-n+j)\right][n-j]\right\}}{\prod_{j=0}^{n-1}\left\{\left[\prod_{k=a_j+1}^{a_{j+1}-1}(N-k+1)\right][N-a_{j+1}+1]\right\}} \\
&= \frac{A}{B}.
\end{aligned}
$$

Let's look at the numerator A and denominator B separately.

$$
\begin{aligned}
B &= \prod_{j=1}^{n-1}\left\{\left[\prod_{k=a_j+1}^{a_{j+1}-1}(N-k+1)\right][N-a_{j+1}+1]\right\} \\
&= \prod_{j=1}^{n-1}\prod_{k=a_j+1}^{a_{j+1}}(N-k+1)
\end{aligned}
$$

$$= \prod_{k=1}^{a_n} (N-k+1)$$
$$= \frac{N!}{(N-a_n)!}.$$

For the numerator A, note that $\prod_{j=0}^{n-1} [n-j] = n!$. We thus have:

$$\begin{aligned}
A &= \prod_{j=0}^{n-1} \left\{ \left[\prod_{k=a_j+1}^{a_{j+1}-1} (N-k+1-n+j) \right] [n-j] \right\} \\
&= n! \prod_{j=0}^{n-1} \prod_{k=a_j+1}^{a_{j+1}-1} (N-k+1-n+j) \\
&= n! \prod_{j=0}^{n-1} \frac{[N-(a_j+1)+1-n+j]!}{[N-a_{j+1}+1-n+j]!} \\
&= n! \; \frac{[N-(a_0+1)+1-n]!}{[N-a_1+1-n]!} \\
&\qquad \times \frac{[N-(a_1+1)+1-n+1]!}{[N-a_2+1-n+1]!} \\
&\qquad \times \frac{[N-(a_2+1)+1-n+2]!}{[N-a_3+1-n+2]!} \\
&\qquad \times \ldots \times \frac{[N-(a_{n-1})+1-n+(n-1)]!}{[N-a_n+1-n+(n-1)]!} \\
&= \frac{n!\,(N-n)!}{(N-a_n)!}.
\end{aligned}$$

Combining the expressions for A and B, we have

$$P(\mathcal{S}) = \frac{n!\,(N-n)!}{(N-a_n)!} \times \frac{(N-a_n)!}{N!} = \frac{n!\,(N-n)!}{N!}.$$

2.33 We use induction. Clearly, $\mathcal{S}_n$ 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 $\mathcal{U}_{k-1} = \{1, 2, \ldots, k-1\}$, where $k \geq n+1$. We wish to show that $\mathcal{S}_k$ is an SRS of size n from $\mathcal{U}_k = \{1, 2, \ldots, k\}$. Since $\mathcal{S}_{k-1}$ is an SRS, we know that

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

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 $\mathcal{U}_k$. If $\mathcal{A}$ does not contain unit k, then $\mathcal{A}$ can be

achieved as a sample at step $k-1$ and

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

If $\mathcal{A}$ does contain unit k, then the sample at step $k-1$ must consist of $\mathcal{A}_{k-1} = \mathcal{A} - \{k\}$ plus one unit randomly selected from $\mathcal{U}_{k-1} \cap \mathcal{A}_{k-1}^C$, the set of units in $\mathcal{U}_{k-1}$ that are not in $\mathcal{A}_{k-1}$. Using the independence of U_k and V_k,

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

Thus the probability of any sample at step k is that of an SRS.

2.34 Clopper-Pearson intervals can be calculated in a spreadsheet or in a statistical software package.

Part	n	$\hat{p}$	Normal LCL	Normal UCL	CP LCL	CP UCL
(i)	10	0.5	0.190097	0.809903	0.187086	0.812914
(ii)	10	0.9	0.714058	1.085942	0.554984	0.997471
(iii)	100	0.5	0.402000	0.598000	0.398321	0.601679
(iv)	100	0.9	0.841200	0.958800	0.823777	0.950995
(v)	100	0.99	0.970498	1.009502	0.945541	0.999747
(vi)	1000	0.99	0.983833	0.996167	0.981687	0.995194
(vii)	1000	0.999	0.997041	1.000959	0.994441	0.999975

In R, the function binom.test gives Clopper-Pearson intervals. Here is the function I used:

```
nvec <- c(10, 10, 100, 100, 100, 1000, 1000)
pvec <- c(0.5, 0.9, 0.5, 0.9, 0.99, 0.99, 0.999)

cp <- function (nvec,pvec) {
   result = matrix(0,nrow=length(nvec),ncol=6)
   colnames(result) = c("n","p","lcl.norm","ucl.norm","lcl.bt","ucl.bt")
   result[,1] = nvec
   result[,2] = pvec
   senorm = sqrt(pvec*(1-pvec)/nvec)
   result[,3] = pvec - qnorm(.975)*senorm
```

```
   result[,4] = pvec + qnorm(.975)*senorm
   for (i in 1:length(nvec)) {
      binomtestci = binom.test(nvec[i]*pvec[i],nvec[i])$conf.int
      result[i,5:6] = as.vector(binomtestci)
   }
   result
}

cp(nvec,pvec)
```

In SAS software, the following code can be used (this code is for part (iv)):

```
data cp;
   input group count wgt;
   datalines;
   1 90 1000
   0 10 1000
;

data cp_long;
   set cp;
   do i = 1 to count;
      output;
   end;
run;

proc surveyfreq data=cp_long;
  tables group/ cl df=10000;  /* To use the t distribution instead of normal, omit the df
     option */
  weight wgt;
  title 'Normal Approximation CI';
run;

proc surveyfreq data=cp_long;
  tables group/ cl (type = clopperpearson ) ;
  weight wgt;
  title 'Clopper-Pearson CI';
run;
```

(c) The normal interval for these parts has upper limit greater than 1, so it is including impossible values.

2.35 I always use this activity in undergraduate 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. They also tend to pick rectangles that give a value of s^2 that is too small, i.e., their sample is not spread out enough.

This is also a good activity for reviewing confidence intervals and other concepts from an introductory statistics class. Students also ask questions about what if the same random number comes up more than once.

Chapter 3

Stratified Sampling

3.1 Answers will vary.

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)

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

3.3 (a) $\bar{y}_{\mathcal{U}} = 71.83333$, $S^2 = 86.16667$.

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

(d) $\begin{pmatrix} 3 \\ 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 9.$

(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;
   length stratum $ 18;
   input stratum $ _total_ returns percfem percagree;
   females = round(returns*percfem/100);
   males = returns - females;
   sampwt = _total_/returns;
   datalines;
Literature 9100 636 38 37
Classics   1950 451 27 23
Philosophy 5500 481 18 23
History    10850 611 19 29
Linguistics     2100 493 36 19
PoliSci    5500 575 13 43
Sociology  9000 588 26 41
;

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 total=acls plots=none mean clm sum clsum;
   stratum stratum;
   weight sampwt;
   var femind;
run;
```

We obtain $\hat{t}_{\text{str}} = 10858$ with SE 313, and $\hat{p}_{\text{str}} = 0.246772$ with SE 0.007121. These values differ from those in Example 3.4 because of rounding.

We need the stratification information to calculate the SEs. The weights by themselves are not sufficient for that purpose.

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

(b)

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

$$\begin{aligned}
\text{SE}\left[\hat{p}_{\text{str}}\right] &= \sqrt{\sum_{h=1}^{7} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}} \\
&= \sqrt{1.46 \times 10^{-5} + 5.94 \times 10^{-7} + \cdots + 1.61 \times 10^{-5}} \\
&= 0.0079.
\end{aligned}$$

Here is R code (this does the same type of calculations as in the SAS code in the previous exercise):

```
acls <- data.frame(
  discipline = c("Literature", "Classics", "Philosophy", "History", "Linguistics", "
    PoliSci", "Sociology"),
  poptotal = c( 9100, 1950, 5500, 10850, 2100, 5500, 9000),
  returns =c(636, 451, 481, 611, 493, 575, 588),
  percfem =c(38, 27, 18, 19, 36, 13, 26),
  percagree =c(37, 23, 23, 29, 19, 43, 41))
acls$numagree <- with(acls,round(returns*percagree/100))
# create data frame
aclsag <- acls[rep(1:nrow(acls),times=acls$numagree),]
```

```
aclsag$agree <- 1
aclsdisag <- acls[rep(1:nrow(acls),times=(acls$returns - acls$numagree)),]
aclsdisag$agree <- 0
aclsdf <- rbind(aclsag,aclsdisag)
dacls <- svydesign(id=~1,strata=~discipline,fpc=~poptotal,data=aclsdf)
svymean(~agree,dacls)
##
##          mean     SE
## agree 0.33355 0.0079
```

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,

$$\begin{aligned} V_{\text{str}}[\hat{p}_{\text{str}}] &= \left(\frac{350}{900}\right)^2 \frac{(0.45)(0.55)}{350} + \left(\frac{450}{900}\right)^2 \frac{(0.25)(0.75)}{450} + \left(\frac{100}{900}\right)^2 \frac{(0.03)(0.97)}{100} \\ &= 0.000215. \end{aligned}$$

If these proportions hold in the population, then

$$p = \frac{35}{90}(0.45) + \frac{45}{90}(0.25) + \frac{10}{90}(0.03) = 0.3033$$

and, with an SRS of size 900,

$$V_{\text{srs}}[\hat{p}_{\text{srs}}] = \frac{(0.3033)(1 - 0.3033)}{900} = 0.000235.$$

The gain in efficiency is given by

$$\frac{V_{\text{str}}[\hat{p}_{\text{str}}]}{V_{\text{srs}}[\hat{p}_{\text{srs}}]} = \frac{0.000215}{0.000235} = 0.9144.$$

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

$$V_{\text{str}}[\hat{p}_{\text{str}}] = \frac{0.193233}{n} \quad \text{and}$$

$$V_{\text{srs}}[\hat{p}_{\text{srs}}] = \frac{0.211322}{n}.$$

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 Chapter 7.

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 2.6 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

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

3.8 Note that in the third edition this exercise was updated with proportionately higher costs (for many applications, these costs are still too low). The allocations are the same as in the second edition, however.

(a) A total of 150,000/300 = 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	100	0.9	0.090
Nonphone	400	0.1	0.005
Total		1.0	0.095

The calculations in the table imply that

$$n_{\text{phone}} = \frac{0.090}{0.095} n;$$

the cost constraints imply that

$$100 n_{\text{phone}} + 400 n_{\text{non}} = 100 n_{\text{phone}} + 400(n - n_{\text{phone}}) = 150{,}000.$$

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 Answers will vary, depending on the sample drawn. The CI for total number of circles has a width of zero because that is one of the stratification variables. In the circle stratum, all of the objects are circles; in the square stratum, none of the objects are circles. The variance for the indicator variable for whether an object is a circle is zero for both strata.

3.10 Answers will vary.

3.11 This will be a stratified random sample if the tickets are placed in the kleroterion slots in random order. It is a stratified random sample of the citizens who are willing to serve, however, not a stratified random sample of all Athenian citizens. And it certainly was not a representative sample of Athenian adults. Citizenship in Athens was restricted to men who had been born free in Athens; it excluded women, foreigners living in Athens, and slaves. And the citizens had to have sufficient leisure to spend their days hanging around the Agora, which might have lessened the participation of poorer citizens. (Many, however, had ample leisure because women ran the households and slaves did most of the other work.)

The strata are the tribes (or, for some elections, subdivisions of tribes were used). The probability that citizen i from tribe h is selected is J/N_h, where N_h is the number of citizens whose tickets are placed in the basket for tribe h. The sample will be self-weighting when $N_1 = N_2 = \ldots = N_{10}$.

The Athenians put in a lot of safeguards to ensure a fair selection process. The procedure was conducted in public. The ticket-inserters were randomly selected on the day the jury was assembled, making it difficult for someone to fix the jury in advance. Of course, an individual ticket-inserter who did not want Euclid on the jury could palm Euclid's ticket and insert it at the bottom of the column where, unless this was the shortest column, it could not be selected. But if the ticket-inserters stirred the baskets, thereby randomizing the order of the tickets, each person present from a particular tribe had an equal chance of being selected, even though the selection probabilities varied across the tribes.

See `https://www.sharonlohr.com/blog/2019/11/20/stratified-random-sampling-aristotle-and-democracy` for a more detailed discussion of the procedure and pictures of kleroteria. Also see Demont (2010) for more information about kleroteria.

3.12 (a) The sample is approximately self-weighting; proportional allocation was used. The weights are:
AJPH: 280/100 = 2.8
AJPM: 103/38 = 2.71053
PM: 164/60 = 2.73333

(b) An fpc should be used when it is desired to estimate characteristics of the 547 articles in the finite population. If you are interested in relationships among variables, or want to use the population as representative of some larger group, don't use the fpc.

(c) Note that you have to define the sampling weights in the software package you're using before doing the calculations. I used variable *sampwt* in the following code. Data set *health_tot* contains the information needed to calculate the fpc.

```
data health_tot;
   length journal $ 4;
   input journal $ _total_ sampsize;
   sampwt = _total_/sampsize;
   datalines;
AJPH 280 100
AJPM 103  38
PM   164  60
;

proc sort data=healthjournals;
   by journal;

data healthj1; /* Merge the sampling weights with the data set */
   merge healthjournals health_tot;
   by journal;
   confhyp = cat("CI ",ConfInt," HT ",HypTest); /*Define new variable with 4 categories*/
   randcat = cat("Sel ",RandomSel," Assn ",RandomAssn);

proc surveymeans data=healthj1 total = health_tot mean clm sum clsum;
   class confhyp randcat;
   weight sampwt;
   strata journal;
   var confhyp randcat NumAuthors;
run;
```

The following table gives the estimated proportions and totals for parts (c) and (d).

Statistics										
Variable	Level	Label	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum	95% CL for Sum	
NumAuthors		NumAuthors	5.437352	0.164555	5.11281694	5.76188719	2974.231579	90.011331	2796.71087	3151.75229
inference	Both		0.632323	0.026592	0.57987804	0.68476804	345.880702	14.545866	317.19329	374.56812
	CI only		0.085756	0.016000	0.05420191	0.11731097	46.908772	8.751797	29.64844	64.16910
	Neither		0.065688	0.014164	0.03775326	0.09362362	35.931579	7.747953	20.65104	51.21212
	PV only		0.216232	0.022644	0.17157252	0.26089164	118.278947	12.386536	93.85017	142.70772
design	Both		0.030184	0.009733	0.01098775	0.04937981	16.510526	5.324109	6.01030	27.01076
	Neither		0.631951	0.027410	0.57789229	0.68600970	345.677193	14.993435	316.10708	375.24731
	RA only		0.075275	0.014895	0.04589903	0.10465102	41.175439	8.147569	25.10677	57.24411
	RS only		0.262590	0.025021	0.21324358	0.31193683	143.636842	13.686517	116.64424	170.62945

(e) The statistics here, calculated with the fpc, apply only to the 547 articles from which the sample was drawn. Still, these were articles in the premier public health journals, and it is somewhat disturbing that 63% used neither random selection nor random assignment.

(f) These statistics apply only to the population of articles from which the sample was drawn. Note that different populations of articles, even from the same set of journals, may well be subject to different editorial practices.

3.13 (a) The sample is about 10% of the population, and sampling fractions in the strata for winners are even larger. Using an fpc will reduce the variances of estimates for this example. The question is whether one wants to use it. Here, I would say yes, because the quantity of interest for (b) is the percentage of books with female authors among the 655 books in the population. If we looked at all 655 books we would know this percentage exactly, with standard error 0. Thus I would say to use the fpcs.

(b) Each book in the sample has a single author, so the estimated percentage with female authors is the estimated percentage where *authorgender* = F. Using weight *p1weight* and the stratification in *stratum*, if we don't use an fpc this estimated percentage is 31.6% with 95% CI [18.98, 44.22].

If we include fpcs for the strata the 95% CI is [19.50 43.70], a little bit smaller. Note that neither CI contains 50%, giving evidence that fewer than half of nominees are women. The Clopper-Pearson CI is similar, [19.6870 45.5954].

(c) To estimate the percentage with at least one female detective, we need to define a new variable that takes value 1 if the book has at least one female detective and 0 otherwise. The estimated percentage is 26.1% with 95% CI [14.03, 38.26].

(d) The estimated total number of female detectives is 176.25 with 95% CI [95.36, 257.14].

3.14 (a) The sampling weight for each observation is calculated as N_h/n_h. SAS software code for computing the weights is as follows:

```
data asafellow;
   set asafellow;
   sampwt = popsize/sampsize;
run;
```

(b) Using the SURVEYMEANS procedure, where *poptotal* is a data set containing the N_hs for each stratum, we have

```
proc surveymeans data=asafellow total=poptotal;
   class field;
   weight sampwt;
   strata awardyr gender;
   var field;
run;
```

This gives an estimated proportion of 0.760 for members in academia, with 95% CI [0.684 0.83]. This is much higher than the proportion of ASA members in academia.

(c) There are missing data for the variable *math*. The method in Example 3.4 for the missing data assumes that the missing values are like the non-missing values in the stratum. We take N_h to be the population count in the stratum, and n_h the number of observations for which *math* is not missing. The following table gives the calculations.

Year	Gender	N_h	n_h	$\hat{p}_h$	s_h^2	$N_h\hat{p}_h/N$	Variance term
2010	F	17	2	1	0	0.032136106	0
	M	36	6	0.8333333	0.1666667	0.056710773	0.000107204
2011	F	11	2	0	0	0	0
	M	46	7	0.8571429	0.1428571	0.074534165	0.000130832
2012	F	13	2	0.5	0.5	0.012287335	0.000127751
	M	35	5	0.8	0.2	0.052930057	0.000150085
2013	F	20	3	1	0	0.037807183	0
	M	39	6	0.8333333	0.1666667	0.061436671	0.000127751
2014	F	21	4	0.25	0.25	0.009924386	7.97328E-05
	M	42	7	0.8571429	0.1428571	0.068052933	0.000107204
2015	F	18	2	1	0	0.034026465	0
	M	44	8	0.875	0.125	0.072778828	8.84431E-05
2016	F	21	4	1	0	0.039697543	0
	M	44	5	0.8	0.2	0.066540643	0.000245282
2017	F	19	2	1	0	0.035916824	0
	M	43	9	0.3333333	0.25	0.027095145	0.000145122
2018	F	20	4	1	0	0.037807183	0
	M	40	6	0.5	0.3	0.037807183	0.000242995
Total		529				0.757489423	0.001552402

We estimate $\hat{p} = 0.76$ with standard error $\sqrt{0.001552402} = 0.039$.

Or, you can do this in statistical software by using weight N_h/n_{hobs} for each stratum, where n_{hobs} is the number of observations with data for *math*.

3.15 (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

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

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

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

(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

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

(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

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

(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

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

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;
```

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

3.17 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, and 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),

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

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;
```

```
                          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

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

3.18

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.

3.19 (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.

(c) Because strata are sampled independently, we can compare estimates using ANOVA or pairwise t tests adjusted by the Bonferroni method.

3.20 (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

$$\text{SE}\,[\hat{t}_{\text{str}}] = \sqrt{5464} = 73.9.$$

Here is SAS software code and output for estimating these quantities:

```
data otters;
   set datalib.otters;
   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=otters total = strattot mean clm sum clsum;
   stratum habitat;
   weight sampwt;
   var holts;
run;
```

```
                              Std Error
Variable            Mean         of Mean       95% CL for Mean
holts           4.154912        0.311903    3.53396136 4.77586336

Variable             Sum         Std Dev       95% CL for Sum
holts        984.714229       73.920990   837.548842 1131.87962
```

3.21 (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 ≤ 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:

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

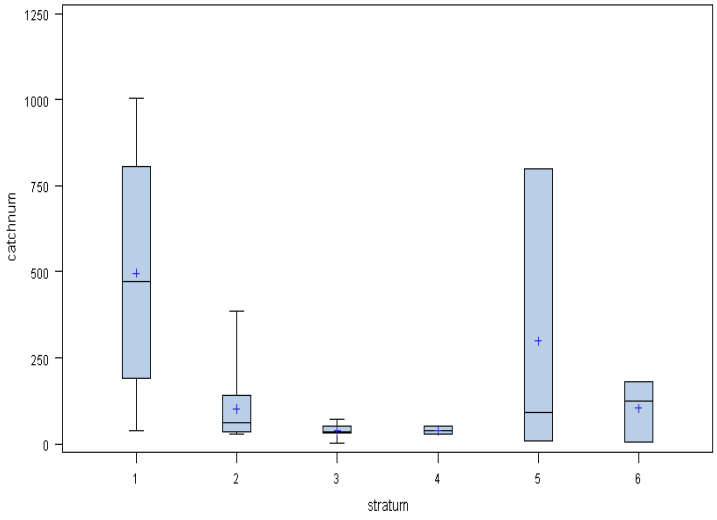
and

$$\begin{aligned}\mathrm{SE}(\hat{p}_{\mathrm{str}}) &= \sqrt{(7.19\times10^{-8})+0+(5.80\times10^{-9})+\cdots+(2.83\times10^{-8})+0} \\ &= 9.75\times10^{-4}.\end{aligned}$$

A 95% confidence interval is

$$0.1566 \pm 1.96(9.75\times10^{-4}) = [0.155, 0.158].$$

3.22 (a) The plot shows a large amount of variation among the strata.



(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

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

equivalently, we may define a new column

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

and calculate

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

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 (3.3) 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 software 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 ;
   weight relwt;
   stratum stratum;
   var catchnum catchwt;
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

3.23 (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	1.01×10^{-4}
2	1960	514	.893	.475	3.90×10^{-5}
3	252	86	.872	.060	4.05×10^{-6}
4	95	67	.866	.022	3.46×10^{-7}
Total	3681	899		.826	1.44×10^{-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.

3.24 (a) Certainly nonresponse and undercoverage are possible sources of selection bias. This survey was done in 1987, when response rates (and landline coverage) were much higher than they are in 2020.

(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].

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

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

3.26 Answers will vary since students select different samples.

3.27 (a) Under proportional allocation, we would take 36.8% (= $100 \times 5803/15769$) of the 1100

observations in the women stratum, and 63.2% of the observations in the men stratum. This would result in a sample size of 405 women and 695 men.

(b) Under Neyman allocation, the sample size for women should be

$$n_W = 1100\frac{5803\sqrt{11}}{5803\sqrt{11} + 9966\sqrt{5}} = 510$$

and the sample size for men should be

$$n_M = 1100\frac{9966\sqrt{5}}{5803\sqrt{11} + 9966\sqrt{5}} = 590.$$

3.28 Table SM3.1 shows the calculations used for this exercise. The sample sizes for proportional allocation are in the column $n_h$prop and those for Neyman allocation are in column n_hN. The anticipated variance for the mean of *ugds* for proportional allocation is 98779 and 65299 for Neyman allocation. The allocation in Example 3.12 appears to be better for quantities correlated with *ugds*.

Table 3.1: Stratum population sizes, variances, and allocations.

control	ccbasic	N_h	N_h/N	$n_h$prop	var, prop	S_h^2	N_hS_h	n_hN	var, Neyman
1	15	93	0.0678	14	22914	82200957	843182	30	8527
1	16	85	0.0620	12	12964	47197758	583955	21	6495
1	17	28	0.0204	4	2031	22751656	133556	5	1557
1	18	149	0.1086	22	18227	39890473	941068	34	10680
1	19	51	0.0372	7	1256	7374422	138495	5	1838
1	20	33	0.0241	5	2193	22344007	155989	6	1762
1	21	18	0.0131	3	62	1297710	20505	1	211
1	22	43	0.0313	6	778	5521803	101044	4	1230
2	15	37	0.0270	5	3189	25348824	186286	7	2136
2	16	40	0.0292	6	3151	26178665	204660	7	2622
2	17	98	0.0714	14	6866	21980420	459456	16	5865
2	18	150	0.1093	22	19925	42976655	983349	35	11253
2	19	114	0.0831	17	426	1234158	126646	5	1629
2	20	70	0.0510	10	144	643700	56162	2	814
2	21	200	0.1458	29	347	554667	148952	5	2298
2	22	163	0.1188	24	4305	8583751	477558	17	6383
Sum		1372		200	98779		5560863	200	65299

3.29 (a) Table SM3.2 shows the calculations to obtain the estimates. Column $N_h \times \bar{y}_h$ tells that the estimated population total is $\hat{t}_{str} = 496$. The last column gives

$$\sum_{h=1}^{8}(1 - n_h/N_h)N_h^2 s_h^2/n_h = 5011.643$$

so $SE(\hat{t}_{str}) = \sqrt{5011.643} = 70.792$. Thus, an approximate 95% CI for the population total is

$$[5011.643 - 1.99(70.792), 5011.643 + 1.99(70.792)] = [355.4, 636.6].$$

Table 3.2: Calculations for estimating the population total of Exercise 3.29

Stratum	division	density	N_h	n_h	$\bar{y}_h$	s_h^2	$N_h \times \bar{y}_h$	$N_h^2 \frac{s_h^2}{n_h}$	$\left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}$
1	1	High	12	12	2.167	3.424	26	41.091	0
2	2	High	8	8	3.500	2.571	28	20.571	0
3	3	High	4	4	5.750	0.250	23	1.000	0
4	4	High	16	16	9.188	12.829	147	205.267	0
5	1	Low	384	24	0.375	0.332	144	2036.870	1909.565
6	2	Low	192	12	0.167	0.152	32	465.455	436.364
7	3	Low	240	15	0.133	0.124	32	475.429	445.714
8	4	Low	144	9	0.444	1.028	64	2368.000	2220.000
Total			1000	100			496	5613.682	5011.643

I used 92 (= 100 − 8) df for the t critical value of 1.98, but it would be fine to use the normal distribution critical value of 1.96.

Alternatively, one can calculate these quantities directly from the data file with the SURVEYMEANS procedure, after entering the survey weights in variable SamplingWeight.

```
proc surveymeans data=PITcount mean sum clsum total = props;
   var y;
   weight SamplingWeight;
   strata strat;
   title 'Surveymeans for PIT sample';
run;
```

This gives the following output:

```
                        Std Error                    Std Error
Variable        Mean       of Mean          Sum        of Sum       95% CL for Sum
y           0.496000      0.070793   496.000000     70.792960   355.399071 636.600929
```

(b) The fpc reduces the variance from 5614 to 5012. The high-density strata contribute nothing to the variance because all areas in those strata are sampled.

(c) There is almost certainly measurement error because the observed values of y_i in each area may underestimate the true number of people experiencing homelessness in the area on the night of the count. The volunteer teams may not find every unsheltered person in an area. In fact, in most parts of the country, the teams are instructed to avoid abandoned buildings and other locations that might be unsafe to enter. A team may mistakenly classify a person encountered as homeless when they are not homeless, and vice versa.

Table 3.3: Calculations for estimating the allocations in Exercise 3.30

Stratum	Division	Density	N_h	s_h^2	$N_h s_h$	$\frac{nN_h}{N}$	$n\frac{N_h s_h}{\sum_j N_j s_j}$
1	1	High	12	3.42	22.21	1.20	3.58
2	2	High	8	2.57	12.83	0.80	2.07
3	3	High	4	0.25	2.00	0.40	0.32
4	4	High	16	12.83	57.31	1.60	9.23
5	1	Low	384	0.33	221.10	38.40	35.63
6	2	Low	192	0.15	74.74	19.20	12.04
7	3	Low	240	0.12	84.45	24.00	13.61
8	4	Low	144	1.03	145.99	14.40	23.52
Total			1000		620.61	100	100

Stratum	Division	Density	Design 1	Design 2	Design 3	Design 4
1	1	High	2	4	12	12
2	2	High	2	2	8	8
3	3	High	2	2	4	4
4	4	High	2	9	16	16
5	1	Low	37	35	25	2
6	2	Low	18	12	8	2
7	3	Low	23	13	10	2
8	4	Low	14	23	17	54
Total			100	100	100	100

3.30 (a) Table SM3.3 gives the calculations for the four allocations. The formulas for proportional and Neyman allocation are given in the columns headed by $\frac{nN_h}{N}$ and $n\frac{N_h s_h}{\sum_j N_j s_j}$. But the actual allocation must have integer values and a minimum of two observations in each stratum, so the allocations for Designs 1 and 2 are adjusted to meet those restrictions.

(b) The anticipated variance is calculated as $\sum_{h=1}^{8}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{s_h^2}{n_h}$ for each allocation, and the anticipated number of persons counted is calculated as $\sum_{h=1}^{8} n_h\bar{y}_h$. Table SM3.4 gives the calculations and values for the four designs.

(c) The Neyman allocation has the lowest anticipated variance (which one would expect since it is supposed to be optimal, even though it does not account for the fpc). Design 4 has the highest anticipated number of persons identified in the sample (and is optimal for maximizing the number of persons found, if the sample means have the same ordering as the population means), but has by far the highest variance for estimating the total number of persons experiencing homelessness because the sample size in three of the low-density strata is only two.

Design 3, that used by New York City, is not optimal for either goal but its variance is relatively low (and certainly lower than proportional allocation) and it captures nearly as many persons in

Table 3.4: Calculations for the anticipated variance and number of persons in sample for Exercise 3.30

	Anticipated Variance				Anticipated Number of Persons in Sample			
Stratum	Design 1	Design 2	Design 3	Design 4	Design 1	Design 2	Design 3	Design 4
1	205.45	82.18	0	0	4.33	8.67	26.00	26.00
2	61.71	61.71	0	0	7.00	7.00	28.00	28.00
3	1.00	1.00	0	0	11.50	11.50	23.00	23.00
4	1436.87	159.65	0	0	18.38	82.69	147.00	147.00
5	1193.91	1269.41	1828.09	24315.13	13.88	13.13	9.38	0.75
6	281.21	436.36	669.09	2763.64	3.00	2.00	1.33	0.33
7	280.35	518.86	683.43	3536.00	3.07	1.73	1.33	0.27
8	1374.29	778.61	1105.65	246.67	6.22	10.22	7.56	24.00
Total	4834.79	3307.78	4286.26	30861.43	67.37	136.93	243.60	249.35

the unsheltered population as Design 4. It is thus a good choice for meeting both goals.

3.31 The guidelines state that a census is "the most complete and accurate information available" but in many situations that is not true. When a large land area must be covered, a census means that the volunteers can do only a cursory investigation and will likely miss many people. When the census will have a lot of undercoverage or measurement error, a sample will be better.

3.32 Equation (3.13) says that proportional allocation gives smaller variance than an SRS unless

$$\text{SSB} < \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) S_h^2.$$

Thus, a population in which all stratum means are equal (that is, $\bar{y}_{h\mathcal{U}} = \bar{y}_{\mathcal{U}}$ for all h), but the within-stratum variances S_h^2 are positive, will have stratification less efficient than an SRS. In general, though, we strive to have stratifications for which the stratum means are different and hence stratification improves efficiency.

3.33 (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, $p = (.4)(.10) + (.6)(.03) = .058$

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

3.34 (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,

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

3.35 (a) 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. 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

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

for $k = 1, \ldots, H$, 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. To approach the problem from this perspective, 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

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

(b) We have

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) n.$$

If the cost is fixed, we substitute this expression for n_h into the cost function:

$$C - c_0 = \sum_{h=1}^{H} c_h \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) n.$$

Solving for n,

$$\begin{aligned} n &= (C - c_0) \frac{\sum_{l=1}^{H} (N_l S_l)/\sqrt{c_l}}{\sum_{h=1}^{H} c_h \left(\frac{N_h S_h}{\sqrt{c_h}} \right)} \\ &= (C - c_0) \frac{\sum_{l=1}^{H} (N_l S_l)/\sqrt{c_l}}{\sum_{h=1}^{H} \sqrt{c_h} N_h S_h}. \end{aligned}$$

(c) If the variance is fixed at V_0, we substitute the expression for n_h into the variance function in (3.4):

$$\begin{aligned} V_0 &= V(\bar{y}_{\text{str}}) \\ &= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h} \right) \left(\frac{N_h}{N} \right)^2 \frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{H} \frac{N_h S_h^2}{N^2} \end{aligned}$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{nN_hS_h/\sqrt{c_h}} \left(\sum_{l=1}^{H} \frac{N_lS_l}{\sqrt{c_l}}\right) - \sum_{h=1}^{H} \frac{N_h}{N^2} S_h^2$$
$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2\sqrt{c_h}}{nN_hS_h} \left(\sum_{l=1}^{H} \frac{N_lS_l}{\sqrt{c_l}}\right) - \sum_{h=1}^{H} \frac{N_h}{N^2} S_h^2.$$

The last term, $\sum_{h=1}^{H} N_h/N^2S_h^2$, is the effect of the fpc. The expression is simpler if this can be ignored, but let's plow forward, including it. Solving for n, we have:

$$n = \frac{\displaystyle\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2\sqrt{c_h}}{N_hS_h} \left(\sum_{l=1}^{H} \frac{N_lS_l}{\sqrt{c_l}}\right)}{\displaystyle V_0 + \sum_{h=1}^{H} \frac{N_h}{N^2} S_h^2}$$
$$= \frac{\displaystyle\sum_{h=1}^{H} \frac{N_hS_h\sqrt{c_h}}{N^2} \left(\sum_{l=1}^{H} \frac{N_lS_l}{\sqrt{c_l}}\right)}{\displaystyle V_0 + \sum_{h=1}^{H} \frac{N_h}{N^2} S_h^2}$$
$$= \frac{\displaystyle\sum_{h=1}^{H} N_hS_h\sqrt{c_h} \left(\sum_{l=1}^{H} \frac{N_lS_l}{\sqrt{c_l}}\right)}{\displaystyle V_0N^2 + \sum_{h=1}^{H} N_hS_h^2}.$$

3.36 (a) We substitute $n_{h,\text{Neyman}}$ for n_h in (3.3):

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

(b)

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

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

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

3.37 Note that

$$\sum_{k=1}^{K} a_k V(\hat{t}_{y_k}) = \sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{\sum a_k S_{kh}^2}{n_h}.$$

The result follows by substituting $\sum_{k=1}^{K} a_k S_{kh}^2$ for S_h^2 in (3.15).

3.38 (a) The argument is similar to that in Exercise 3.35. In fact, we can just substitute into the previous argument.

Note that (3.15), when $c_h = 1$, gives the optimal allocation when one is minimizing $\sum_{h=1}^{H} N_h^2 V(\bar{y}_h)$ subject to $\sum_{h=1}^{H} n_h = n$. In this problem, we want to minimize $\sum_{h=1}^{H} (X_h^q/\bar{y}_{h\mathcal{U}})^2 V(\bar{y}_h)$ subject to $\sum_{h=1}^{H} n_h = n$. Thus, we can just substitute $X_h^q/\bar{y}_{h\mathcal{U}}$ for N_h in (3.15), keeping $c_h = 1$, to give the allocation that minimizes the weighted sum of the stratum variances.

(b) When $X_h = t_h$ and $q = 1$, $X_h^q/\bar{y}_{h\mathcal{U}} = N_h$, so this is Neyman allocation.

(c) When $q = 0$ and $S_h/\bar{y}_{h\mathcal{U}} = S_l/\bar{y}_{l\mathcal{U}}$, then from part (a),

$$n_h = n\frac{X_h^q S_h/\bar{y}_{h\mathcal{U}}}{\sum_{l=1}^{H} X_l^q S_l/\bar{y}_{l\mathcal{U}}} = n/H.$$

3.39 Table SM3.5 give the fractions n_h/n for each allocation. As q moves from 0 to 1, more of the sample is allocated to the strata with large values of S_h.

Table 3.5: Power allocations for a stratified sample from Example 3.12.

Stratum	N_h	S_h	$\bar{y}_{h\mathcal{U}}$	$q = 0$	$q = 0.25$	$q = 0.5$	$q = 1$
Very small	195	251	639	0.094	0.063	0.040	0.014
Small, primarily nonresidential	45	784	2004	0.093	0.058	0.034	0.010
Small, primarily residential	123	515	1612	0.076	0.057	0.041	0.018
Small, highly residential	347	508	1640	0.074	0.072	0.067	0.049
Medium, primarily nonresidential	80	2490	5927	0.100	0.094	0.083	0.055
Medium, primarily residential	160	2150	5275	0.097	0.105	0.107	0.095
Medium, highly residential	158	1473	4096	0.086	0.087	0.083	0.064
Large, primarily nonresidential	95	11274	21567	0.125	0.168	0.214	0.296
Large, primarily residential	126	9178	18730	0.117	0.163	0.215	0.319
Large, highly residential	43	6844	11772	0.139	0.132	0.118	0.081

3.40 To show the hint, write

$$\begin{aligned}\frac{1}{n_h} &= \frac{1}{k} - \left(\frac{1}{k} - \frac{1}{k+1}\right) - \left(\frac{1}{k+1} - \frac{1}{k+2}\right) - \ldots - \left(\frac{1}{n_h - 1} - \frac{1}{n_h}\right) \\ &= \frac{1}{k} - \frac{1}{(k)(k+1)} - \frac{1}{(k+1)(k+2)} - \frac{1}{(k+2)(k+3)} - \ldots - \frac{1}{(n_h-1)(n_h)}.\end{aligned}$$

We first look at the result for $k \geq 2$. Note that from (3.3),

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

The second term does not depend on n_h, so we want to show this allocation minimizes the first term. From part (a),

$$\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{H} N_h^2 S_h^2 \left[\frac{1}{k} - \frac{1}{(k)(k+1)} - \frac{1}{(k+1)(k+2)} - \frac{1}{(k+2)(k+3)} - \ldots - \frac{1}{(n_h-1)(n_h)}\right]$$

$$= \frac{1}{k}\sum_{h=1}^{H} N_h^2 S_h^2 - \sum_{h=1}^{H} N_h^2 S_h^2 \left[\frac{1}{(k)(k+1)} + \frac{1}{(k+1)(k+2)} + \ldots + \frac{1}{(n_h-1)(n_h)}\right]$$

The first term is a constant, so the variance is minimized when the n_h values are allocated so as to maximize the sum $\sum_{h=1}^{H} N_h^2 S_h^2 \left[\frac{1}{(k)(k+1)} + \frac{1}{(k+1)(k+2)} + \ldots + \frac{1}{(n_h-1)(n_h)}\right]$. This will occur when the $n - kH$ largest terms are chosen from the matrix.

Now, the identity in part (a) is valid only for $n_h \geq 2$. When $k = 1$, each stratum gets one observation. So we only need to look at strata that will get additional observations, for which we will have $n_h \geq 2$. Any strata that do not receive additional observations will not be part of the sum.

3.41 (a) Each time a unit from stratum h appears in the stream, the count N_h is incremented by one. At the end of the process, the algorithm has counted the number of units in each stratum.

(b) If there is one stratum, then steps i and ii reduce to incrementing the population size, and only step iii is used for drawing the sample. Parts a and b of step iii are the same as the procedure in Chapter 2.

3.47 (d) For the 1920 census, equal proportions takes seats away from New York, North Carolina, and Virginia and gives them to Rhode Island, Vermont, and New Mexico.

3.49 (a) Define the variable one to have the value 1 for every observation. Then $\sum_{i\in\mathcal{S}} w_i = N$. Here, $\sum_{i\in\mathcal{S}} w_i = 85174776$. The standard error is zero because this is a stratified random 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.

(b) The estimated total number of truck miles driven is 1.115×10^{12}; the standard error is 6492344384 and a 95% CI is $[1.102\times10^{12}, 1.127\times10^{12}]$.

(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

4.1 Answers will vary.

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

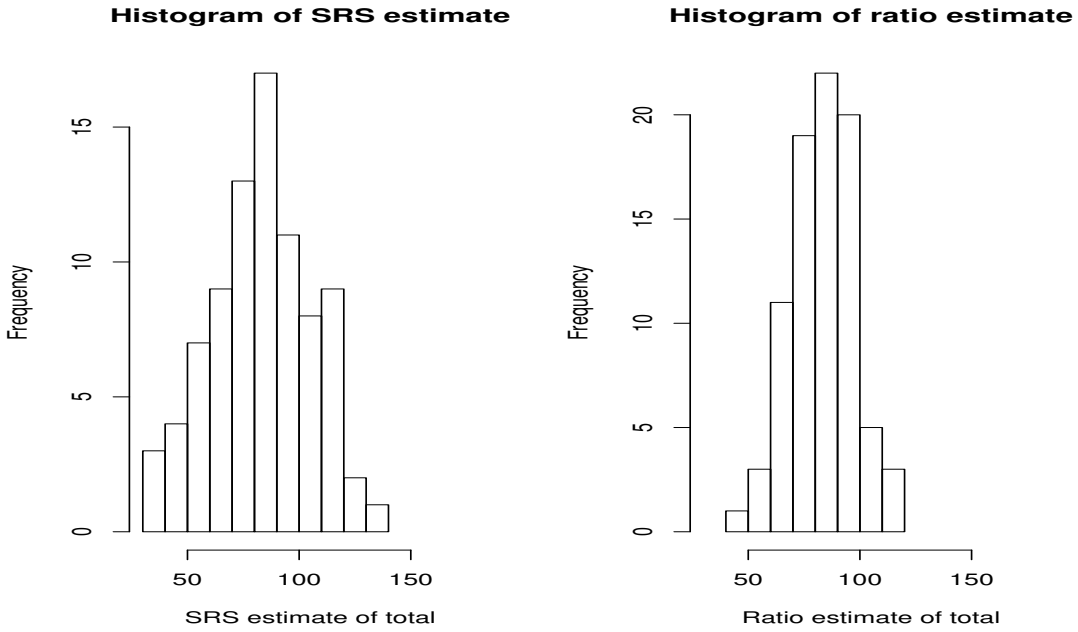
(b)

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

Sample Number	Sample $\mathcal{S}$	$\bar{x}_{\mathcal{S}}$	$\bar{y}_{\mathcal{S}}$	$\hat{B}$	$\hat{t}_{\text{SRS}}$	$\hat{t}_{yr}$
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
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

Sample Number	Sample $\mathcal{S}$	$\bar{x}_{\mathcal{S}}$	$\bar{y}_{\mathcal{S}}$	$\hat{B}$	$\hat{t}_{\text{SRS}}$	$\hat{t}_{yr}$
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)

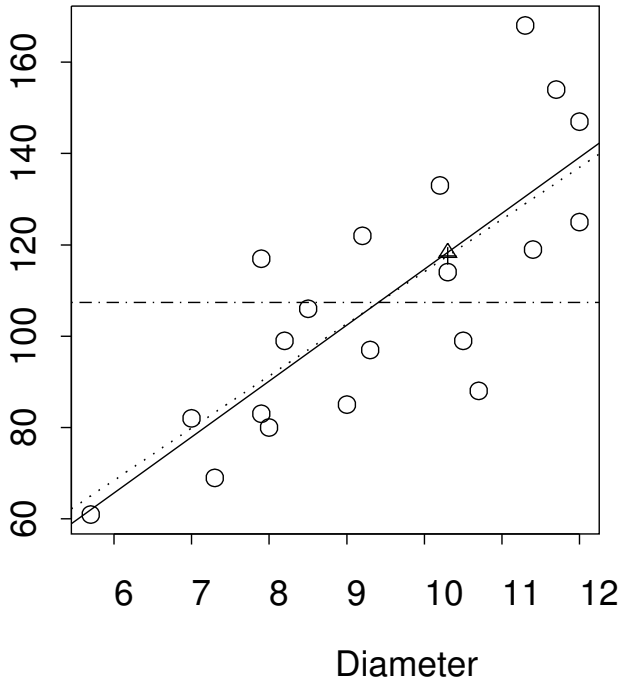


The shapes are actually quite similar; however, it appears that the histogram of the ratio estimator is a little less spread out.

(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 (4.8), 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 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 (4.11), 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 $\text{SE}(\hat{\bar{y}}) = 4.35$. 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 the SURVEYREG procedure in SAS software or using *svyglm* in R are larger since they use $\hat{V}_2$ from Section 11.6, and the two estimates are only asymptotically equivalent.

Here is code and output for SAS software:

```
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;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Lower	Upper
Mean age of trees	118.363449	5.20417420	22.74	<.0001	107.47	129.26

4.4 (a) To estimate the percentage of large-motorboat owners who have children, we can use $\hat{p}_1 = 396/472 = 0.839$. This is a ratio estimator, because the domain sample size (472) is a random variable—if a different sample were taken, it is possible for n_d to be a different number. Ignoring the fpc in (4.22), the standard error is

$$\text{SE}(\hat{p}_1) = \sqrt{\frac{0.839(1-0.839)}{472}} = 0.017$$

The average number of children for the 472 respondents in domain 1 is 1.667373, with sample variance 1.398678. Thus an approximate 95% CI for the average number of children in large-motorboat households is

$$1.667 \pm 1.96\sqrt{\frac{1.398678}{472}} = [1.56, 1.77].$$

(c)To estimate the total number of children in the state whose parents register a large motorboat, define

$$x_i = \begin{cases} 1, & \text{if boat owner } i \text{ owns a large motorboat} \\ 0, & \text{otherwise} \end{cases}$$

and define $u_i = y_i x_i$. Then u_i equals the number of children if boat owner i has a large motorboat, and equals 0 otherwise. The frequency distribution for the variable u is then

Number of Children	Number of Respondents
0	1104
1	139
2	166
3	63
4	19
5	5
6	3
8	1
Total	1500

Now $\bar{u} = 0.52466$ and $s_u^2 = 1.0394178$, so $\hat{t}_{yd} = 400{,}000(0.524666) = 209{,}867$ and

$$\text{SE}\left(\hat{t}_{yd}\right) = \text{SE}\left(\hat{t}_u\right) = \sqrt{\left(1 - \frac{1500}{400{,}000}\right)(400{,}000)^2 \frac{1.0394178}{1500}} = 10{,}509.8$$

The variable u_i counts the number of children in household i that belong to a household with a large open motorboat.

Here is SAS code for estimating the quantities in this problem. The code also gives estimates for domain 2.

```
data boats;
   input boatsize children frequen;
   sampwt = 266.6666667;
   datalines;
1 0 76
1 1 139
1 2 166
1 3 63
1 4 19
1 5 5
1 6 3
1 8 1
2 0 528 /* There are 1028 observations in domain 2, small boats */
2 1 189
2 2 225
```

```
2 3 76
2 4 8
2 5 2
;

/* Expand the data so there are 1500 observations, each representing 1 respondent */
data boats2;
   set boats;
   do i = 1 to frequen;
      if children = 0 then childyes = 0;
      else if children >=1 then childyes = 1;
      output;
   end;
run;

proc freq data = boats2; /* check that data were expanded correctly */
   tables boatsize*children;
run;

proc surveymeans data=boats2 total = 400000 mean sum clm clsum;
   var childyes children;
   weight sampwt;
   domain boatsize;
run;
```

The output is given below.

The SURVEYMEANS Procedure

Statistics for boatsize Domains									
boatsize	Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum	95% CL for Sum	
1	childyes	0.838983	0.016892	0.80584940	0.87211670	105600	4545.526696	96683.732	114516.268
	children	1.667373	0.054295	1.56087149	1.77387427	209867	10510	189251.231	230482.102
2	childyes	0.486381	0.015565	0.45585038	0.51691226	133333	4861.128319	123797.998	142868.669
	children	0.884241	0.032897	0.81971164	0.94877085	242400	9962.867962	222857.358	261942.642

4.5 The weight for each observation is 864/150. The estimated proportion of female members who are in academia is 0.522, with 95% CI [0.376, 0.668]. The estimated total number of female members in academia is 138.24, with 95% CI [86.964, 189.52].

4.6 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,

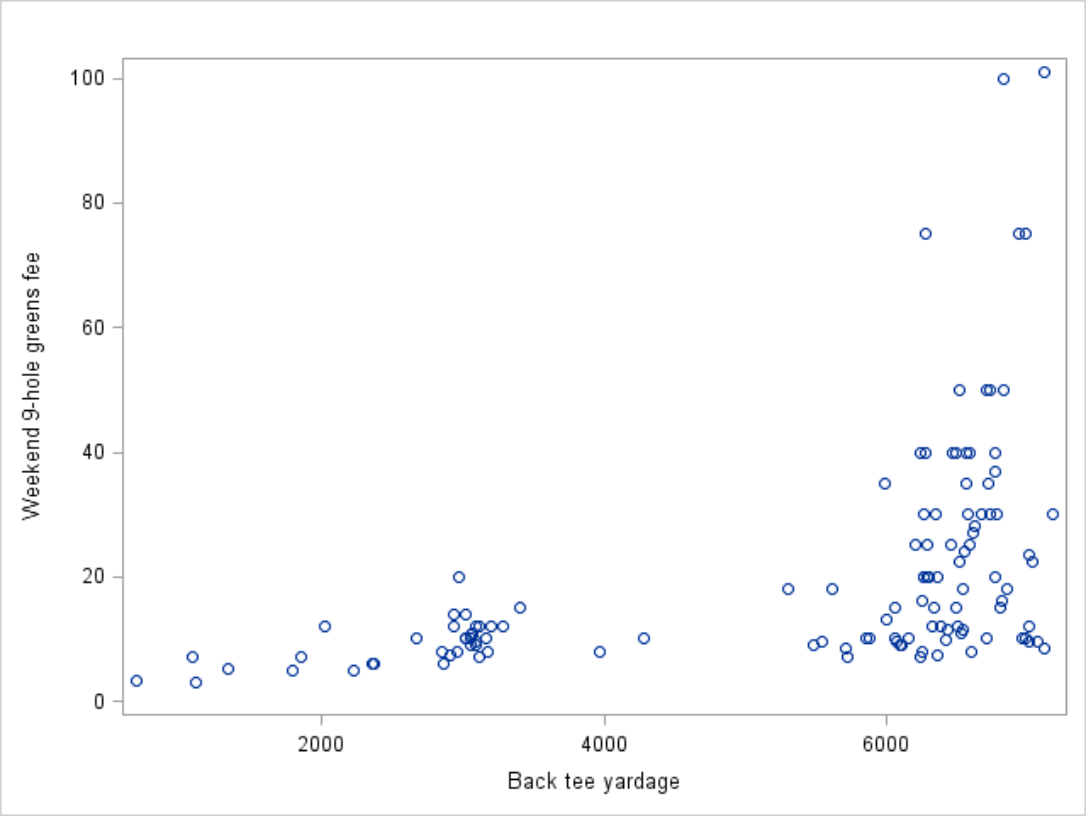
$$\text{SE}[\bar{y}_d] = \sqrt{\frac{395.498}{85}} = 2.16.$$

Here is SAS code:

```
data golfsrs;
   set datalib.golfsrs;
   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;
```

4.7 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

$$\hat{y} = -4.1035032 + 0.0045840x,$$

so that $\hat{\bar{y}}_{reg} = -4.1035032 + 0.0045840 * (5415) = 20.71886$. The sample variance of the residuals is $s_e^2 = 253.35018$. This leads to a standard error (ignoring the fpc) of

$$SE(\hat{\bar{y}}_{reg}) = \sqrt{253.35018/120} = 1.45.$$

Note that statistical survey software, using formula $\hat{V}_2$, gives a standard error of 1.52.

4.8 (a) 88 courses have a golf professional. For these 88 courses, $\bar{y}_{d_1} = 23.5983$ and $s_{d_1}^2 = 387.7194$, 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 $s_{d2}^2 = 19.146$, so

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

4.9 (a) Figure SM4.1 shows the output giving the percentage of arrests for each type of crime. Look under the column "Row Percent" in the row corresponding to "1" for each type. Note that the set of homicides in the data set is too small to reach any conclusions.

(b) The estimates come from the following code. We estimate a total of 76120 aggravated assaults that are domestic-related, with standard error 10304 and 95% CI [55920, 96319].

```
proc surveyfreq data=crimes;
   weight sampwt;
   tables crimetype*domestic / cl clwt;
run;
```

4.10 (a) Figure SM4.2 shows the plot.

(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

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

Using (4.13), then, with farms87 as the auxiliary variable,

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

Figure 4.1: Output for Exercise 4.9(a).

The SAS System

The SURVEYFREQ Procedure

Table of crimetype by arrest

crimetype	arrest	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent		Row Percent	Std Err of Row Percent	95% Confidence Limits for Row Percent	
	1	19	26783	6133	0.3800	0.0870	0.2094	0.5506	10.5556	2.2905	6.0652	15.0459
	Total	180	253732	18570	3.6000	0.2635	3.0835	4.1165	100.0000			
sexualasslt	0	23	32421	6745	0.4600	0.0957	0.2724	0.6476	79.3103	7.5229	64.5621	94.0586
	1	6	8458	3451	0.1200	0.0490	0.0240	0.2160	20.6897	7.5229	5.9414	35.4379
	Total	29	40879	7570	0.5800	0.1074	0.3694	0.7906	100.0000			
simpleasslt	0	798	1124878	36508	15.9600	0.5180	14.9445	16.9755	78.5433	1.2880	76.0182	81.0684
	1	218	307297	20356	4.3600	0.2888	3.7938	4.9262	21.4567	1.2880	18.9316	23.9818
	Total	1016	1432175	40111	20.3200	0.5691	19.2043	21.4357	100.0000			
theft	0	869	1224961	37774	17.3800	0.5360	16.3293	18.4307	87.9555	1.0356	85.9252	89.9857
	1	119	167745	15195	2.3800	0.2156	1.9574	2.8026	12.0445	1.0356	10.0143	14.0748
	Total	988	1392706	39694	19.7600	0.5632	18.6559	20.8641	100.0000			
threat	0	224	315755	20621	4.4800	0.2926	3.9064	5.0536	92.5620	1.6869	89.2550	95.8690
	1	18	25373	5970	0.3600	0.0847	0.1939	0.5261	7.4380	1.6869	4.1310	10.7450
	Total	242	341128	21393	4.8400	0.3035	4.2449	5.4351	100.0000			
trespass	0	43	60614	9205	0.8600	0.1306	0.6040	1.1160	28.8591	3.7124	21.5812	36.1369
	1	106	149420	14360	2.1200	0.2037	1.7206	2.5194	71.1409	3.7124	63.8631	78.4188
	Total	149	210034	16950	2.9800	0.2405	2.5085	3.4515	100.0000			
vandalism	0	531	748509	30712	10.6200	0.4358	9.7657	11.4743	92.1875	1.1183	89.9951	94.3799
	1	45	63433	9414	0.9000	0.1336	0.6381	1.1619	7.8125	1.1183	5.6201	10.0049
	Total	576	811942	31826	11.5200	0.4516	10.6348	12.4052	100.0000			
weapon	0	10	14096	4454	0.2000	0.0632	0.0761	0.3239	18.1818	5.2012	7.9851	28.3785
	1	45	63433	9414	0.9000	0.1336	0.6381	1.1619	81.8182	5.2012	71.6215	92.0149
	Total	55	77529	10397	1.1000	0.1475	0.8108	1.3892	100.0000			
Total	0	3617	5098601	44591	72.3400	0.6327	71.0997	73.5803				
	1	1383	1949506	44591	27.6600	0.6327	26.4197	28.9003				
	Total	5000	7048107	0.01339	100.0000							

(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

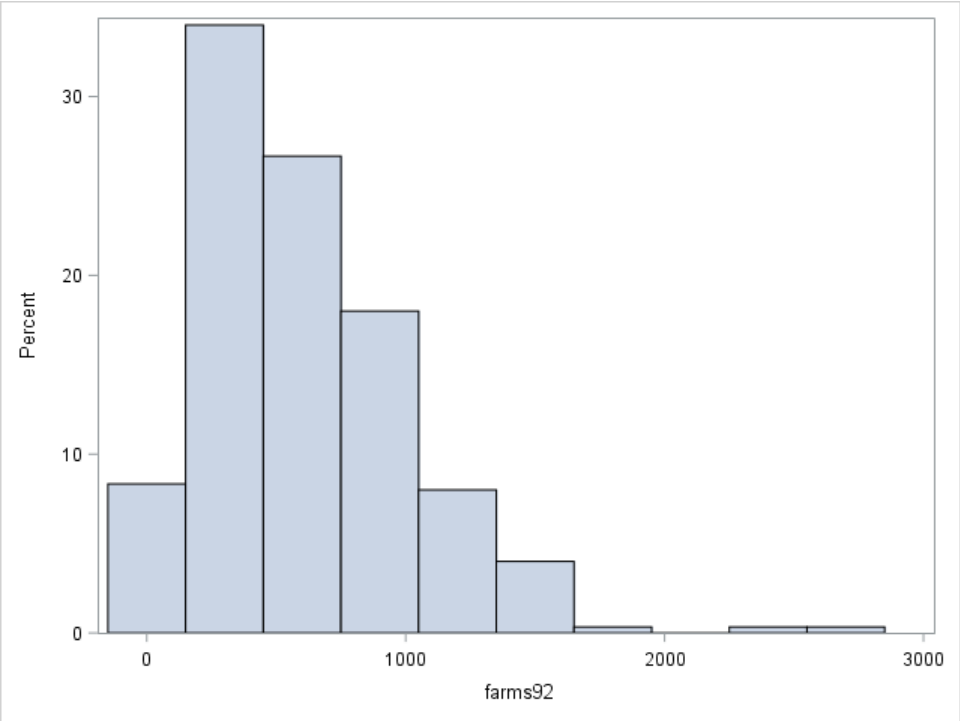
$$\hat{t}_{y\text{reg}} = 3078\hat{\bar{y}}_{\text{reg}} = 916{,}927{,}075.$$

The estimated variance of the residuals from the regression is $s_e^2 = 118{,}293{,}647{,}832$, which implies from (4.20) that

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

(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

Figure 4.2: Plot for Exercise 4.10(a).

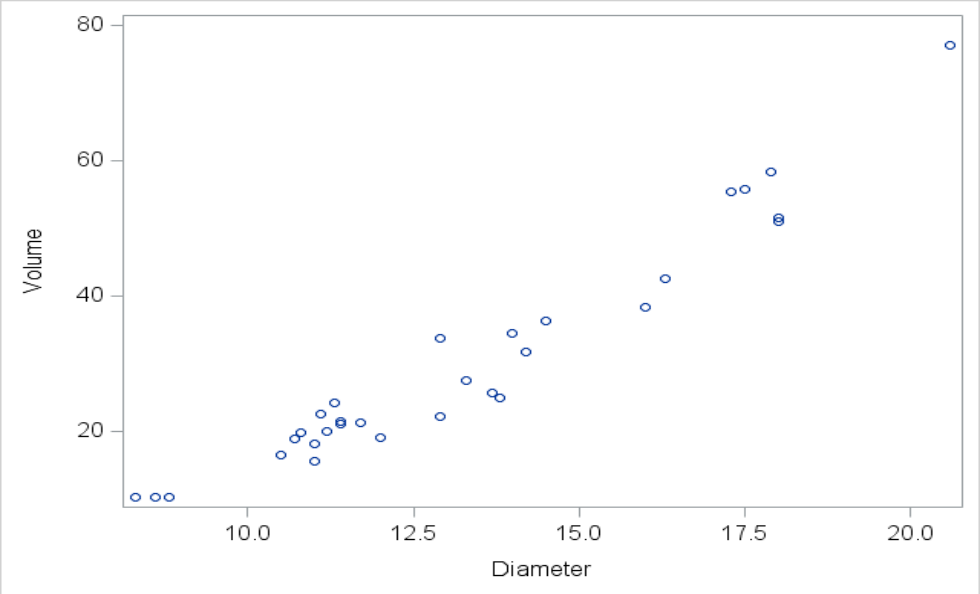


the SRS estimate $N\bar{y}$. The correlation of acres92 and acres87, however, exceeds 0.99. Here are the various estimates for the population total of acres92:

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.

4.11 (a)



(b) Here is code and output from SAS:

```
data cherry;
  set datalib.cherry;
  sampwt = 2967/31;
  xtotal = 41835;
  vol_times_xt = volume*xtotal;
;

proc sgplot data=cherry;
   scatter x = diameter y = volume;
   xaxis label = "Diameter" VALUEATTRS=(Size=12) labelattrs=(size=12);
   yaxis label = "Volume" VALUEATTRS=(Size=12) labelattrs=(size=12);

proc surveymeans data = cherry total=2967 mean clm sum clsum ratio;
   weight sampwt;
   var diameter volume;
   ratio 'vol/diam' volume/diameter;
   ratio 'voltot/diam' vol_times_xt/diameter;
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 volume=diameter / clparm solution;
   weight sampwt;
   estimate 'Total volume' intercept 2967 diameter 41835;
```

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

The ESTIMATE statement gives:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval
Total volume	102318.860	2233.70776	45.81	<.0001	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.

4.12 These are domain estimates, so we need to use the DOMAIN statement in SAS or the subset statement in the R svymeans function.

```
proc surveymeans data=healthj1 total = health_tot mean clm sum clsum;
   weight sampwt;
   strata journal;
   domain hyptest;
   var asterisks;
run;
```

(a) 34.62% of the articles using hypothesis tests are estimated to use asterisks, with 95% CI [28,87, 40.37].

(b) 91.27% of the articles not using random selection or assignment still use inferential methods. A normal-based CI is [87.18, 95.10]; because this percentage is so close to 1 we should perhaps use the Clopper-Pearson interval, which is [84.7839 95.5155].

4.13 (a, c) The variable *physician* has a skewed distribution, as shown in Figure SM4.4.3. The histogram and scatterplot exclude Cook County, Illinois (with y_i = 15,153) for slightly better visibility. There appears to be increasing variance as population increases, so ratio estimation may be more appropriate.

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

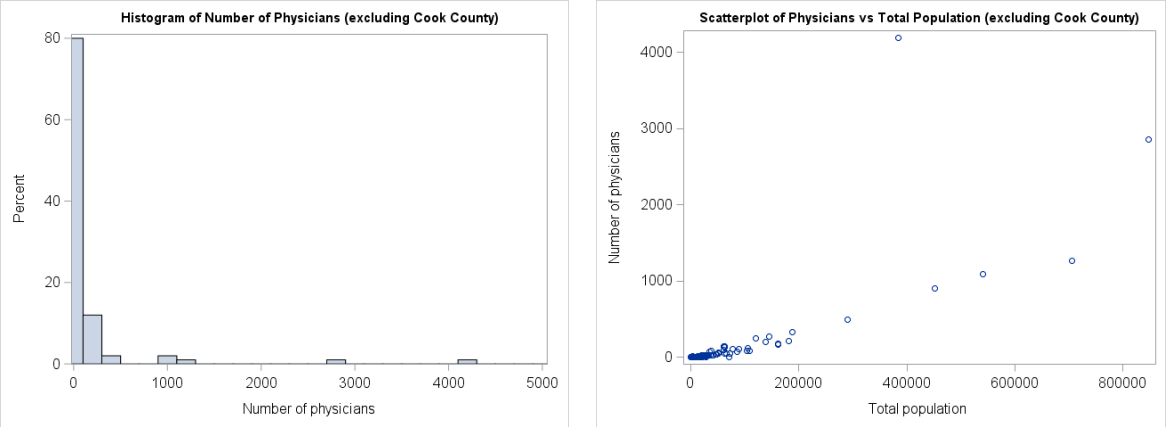
Thus the estimated total number of physicians is

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

with

$$\mathrm{SE}\left(N\bar{y}\right) \quad = \quad N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_y^2}{n}}$$

Figure 4.3: Plots for Exercise 4.13



$$
\begin{aligned}
&= 3141\sqrt{\left(1-\frac{100}{3141}\right)\frac{2{,}534{,}052}{100}} \\
&= 3141\sqrt{24533.75} \\
&= 491{,}983.
\end{aligned}
$$

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 (2.29), 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.

(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 (4.13),

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

Regression estimation:

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

From (4.16) and (4.20),

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

and

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

Consequently,

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

and

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

The standard error from the SURVEYREG procedure is smaller and equals 92,535.

(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 and output for performing the analysis in this and the following two problems:

```
%let poptotal = 255077536;
%let landtotal = 3536278;
%let Ncounties = 3141;
data counties;
   set datalib.counties;
   sampwt = 3141/100;
   phys_x_pop = physician * &poptotal;
   farm_x_land = farmpop * &landtotal;
   veterans_x_pop = veterans * &poptotal;
run;

/* The following histogram is really really skewed because of Cook County */

proc sgplot data=counties;
   histogram physician / binwidth = 500;
   xaxis label = "Number of physicians" min = 0 max = 16000;
```

```
proc sgplot data=counties;
   histogram physician / binwidth = 200;
   xaxis label = "Number of physicians" min = 0 max = 5000 VALUEATTRS=(Size=12)
    labelattrs=(size=12);
   yaxis VALUEATTRS=(Size=12) labelattrs=(size=12);
   title 'Histogram of Number of Physicians (excluding Cook County)';

proc sgplot data=counties;
   where physician < 10000;
  scatter y= physician x=  totpop;
   yaxis label = "Number of physicians" VALUEATTRS=(Size=12) labelattrs=(size=12);
   xaxis label = "Total population " VALUEATTRS=(Size=12) labelattrs=(size=12);
   title 'Scatterplot of Physicians vs Total Population (excluding Cook County)';

proc sgplot data=counties;
   histogram farmpop / binwidth = 200;
   xaxis label = "Farm population" VALUEATTRS=(Size=12) labelattrs=(size=12);
   yaxis VALUEATTRS=(Size=12) labelattrs=(size=12);
   title 'Histogram of Farm Population';
run;

proc sgplot data=counties;
  scatter y= farmpop x=  landarea;
   yaxis label = "Farm population" VALUEATTRS=(Size=12) labelattrs=(size=12);
   xaxis label = "Land area " VALUEATTRS=(Size=12) labelattrs=(size=12);
   title 'Scatterplot of Farm population vs Land area';
run;

proc sgplot data=counties;
   where county ne "Cook";
   histogram veterans / binwidth = 2000;
   xaxis label = "Number of veterans" VALUEATTRS=(Size=12) labelattrs=(size=12);
   yaxis VALUEATTRS=(Size=12) labelattrs=(size=12);
   title 'Histogram of Number of Veterans (excluding Cook County)';

proc sgplot data=counties;
   where county ne "Cook";
  scatter y= veterans x=  totpop;
   yaxis label = "Number of veterans" VALUEATTRS=(Size=12) labelattrs=(size=12);
   xaxis label = "Total population " VALUEATTRS=(Size=12) labelattrs=(size=12);
   title 'Scatterplot of Veterans vs Total Population (excluding Cook County)';
run;

proc surveymeans data=counties total = 3141 mean clm sum clsum;
   weight sampwt;
   var physician farmpop veterans landarea totpop;
   ratio 'physicians' phys_x_pop / totpop;
   ratio 'farmpop' farm_x_land / landarea;
   ratio 'veterans' veterans_x_pop / totpop;
```

```
   ods output Statistics=statsout Ratio=ratioout;
   title 'Ratio estimates';

proc print data=ratioout;

proc print data=statsout;
   title 'Estimates using N ybar';
run;

/* Look at correlations */
proc corr data=counties;
  var physician farmpop veterans landarea totpop;
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;
   title 'Regression estimate: physician';

proc surveyreg data=counties total=3141;
   model physician=totpop / clparm solution ;
   weight sampwt;
   estimate 'Total farm population' intercept 3141  totpop 255077536;
   estimate 'Average farm population' intercept 3141  totpop 255077536/divisor = 3141;
   title 'Regression estimate: farmpop';

proc surveyreg data=counties total=3141;
   model veterans=totpop / clparm solution ;
   weight sampwt;
   estimate 'Total number of veterans' intercept 3141 totpop 255077536;
   estimate 'Average number of veterans' intercept 3141 totpop 255077536/divisor = 3141;
   Title 'Regression estimate: veterans';
run;
```

Here are the estimates of $N\bar{y}$

Obs	VarName	Mean	Lower CLMean	Sum	LowerCLSum
1	physician	297.170000	-13.62293	933411	-42790
2	farmpop	1146.870000	938.53861	3602319	2947950
3	veterans	12250	2961	38476339	9301449
4	landarea	944.920000	718.30722	2967994	2256203

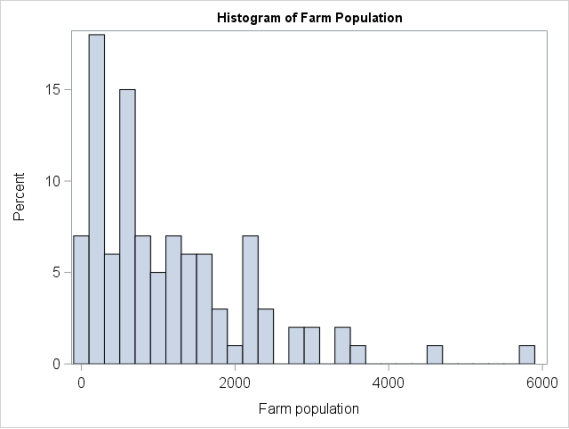
5	totpop	118531	16096	372306374	50558970
6	phys_x_pop	75801391373	-3.4749E9	2.3809217E14	-1.0915E13
7	farm_x_land	4055651150	3318933439	1.27388E13	1.04248E13
8	veterans_x_pop	3.1246258E12	7.55362E11	9.8144498E15	2.37259E15

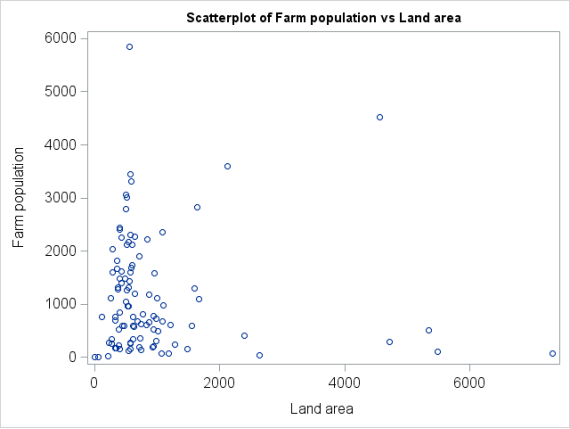
Obs	StdErr	Upper CLMean	StdDev	UpperCLSum
1	156.632533	607.96293	491983	1909612
2	104.994260	1355	329787	4256688
3	4681.145389	21538	14703478	67651229
4	114.207663	1172	358726	3679784
5	51625	220966	162153339	694053777
6	39953440655	1.55078E11	1.2549376E14	4.87099E14
7	371288890	4792368860	1.1662184E12	1.50528E13
8	1.194055E12	5.49389E12	3.7505269E15	1.72563E16

Here are the ratio estimates:

Obs	RatioLabel	NumeratorName	Denominator Name	Ratio	LowerCL	StdErr	UpperCL
1	physicians	phys_x_pop	totpop	639506	465122.6	87885	813889.5
2	farmpop	farm_x_land	landarea	4292058	2965157.7	668727	5618957.7
3	veterans	veterans_x_pop	totpop	26361219	23112963.7	1637046	29609473.9

4.14 (a, c) The distribution appears to be skewed, but not quite as skewed as the distribution in Exercise 4.13.





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

(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

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

(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 (4.16) and (4.20),

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

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

Consequently,

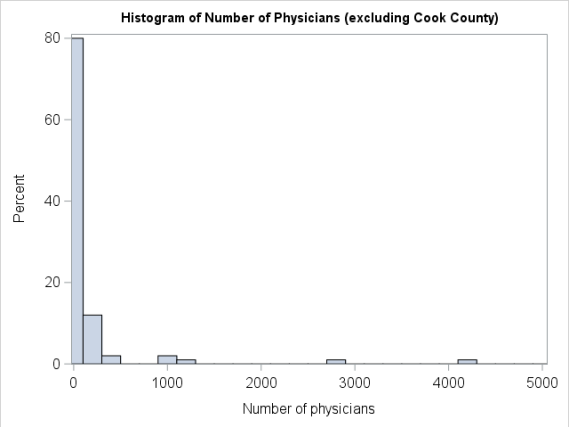
$$\hat{t}_{y\mathrm{reg}} = N\hat{\bar{y}}_{\mathrm{reg}} = 3141(1137.2) = 3{,}571{,}960$$

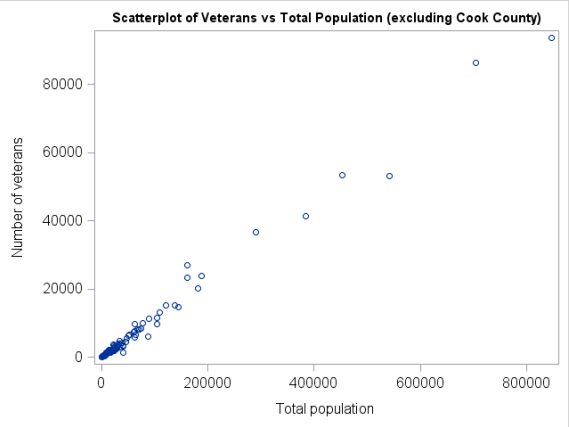
$$\mathrm{SE}\,(\hat{t}_{y\mathrm{reg}}) = N\ \mathrm{SE}\,(\hat{\bar{y}}_{\mathrm{reg}}) = 3141(104.82) = 329{,}230.$$

The SURVEYREG procedure of SAS software gives SE 350,595.

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

4.15 (a,c) As in Exercise 4.13, we omit Cook County, Illinois, from the histogram and scatterplot 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.





The scatterplot appears to be close to a straight line. We would expect ratio or regression estimation to help immensely.

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

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

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

(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 the SURVEYREG procedure is 1,063,699.

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

4.17 (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 (4.26),

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

Not surprisingly, these are the same numbers from the first table. The variances and confidence intervals, however, differ. From (4.27),

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

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:

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

4.18 The ratio of (total male authors)/(total female authors) is 2.16 with 95% CI (including fpc) [0.95, 3.38].

The ratio of (total male detectives)/(total female detectives) is 3.91 with 95% CI (including fpc) [1.64, 6.17]. This CI does not include 1, indicating significantly more male than female detectives.

4.19 An estimated 58.9% of books with female authors had at least one female detective, with 95% CI [34.3, 83.6]. For books with male authors, the corresponding percentage was 11.0%, with 95% CI [0.2, 21.8].

4.20 As $d_i = y_i - Bx_i$, $\bar{d} = \bar{y} - B\bar{x}$. Then, using (A.11),

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

4.23 From (4.8), 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 (4.10),

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

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

4.24 A rigorous proof, showing that the lower order terms are negligible, is beyond the scope of this book. We give an argument for (4.8).

$$\begin{aligned} E[\hat{B} - B] &= E\left[\frac{\bar{y}}{\bar{x}} - \frac{\bar{y}_U}{\bar{x}_U}\right] \\ &= E\left[\frac{\bar{y}}{\bar{x}_U}\left(\frac{\bar{x}_U}{\bar{x}}\right) - \frac{\bar{y}_U}{\bar{x}_U}\right] \\ &= E\left[\frac{\bar{y}}{\bar{x}_U}\left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}}\right) - \frac{\bar{y}_U}{\bar{x}_U}\right] \end{aligned}$$

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

4.25

$$
E\left[\hat{t}_{yr}-t_y\right] = N\mathrm{Bias}(\hat{\bar{y}}_r) \approx N\left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_{\mathcal{U}}}(BS_x^2-RS_xS_y).
$$

Similarly, $\mathrm{Bias}(\hat{B}) = \frac{1}{\bar{x}_{\mathcal{U}}}\mathrm{Bias}(\hat{\bar{y}}_r)$.

4.26 (a) From (4.7),

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

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 $u_iy_i = u_i^2$. Now look at the numerator of (4.35).

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

Consequently,

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

4.27 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

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

4.28 Using (4.10) and (4.18),

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

4.29 From (4.16),

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

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,

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

Since

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

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

4.30

$$\sum_{i=1}^{N}\frac{(y_i - \bar{y}_U - B_1[x_i - \bar{x}_U])^2}{N-1} = \sum_{i=1}^{N}\frac{[(y_i - \bar{y}_U)^2 - 2B_1(x_i - \bar{x}_U)(y_i - \bar{y}_U) + B_1^2(x_i - \bar{x}_U)^2]}{N-1}$$

$$
\begin{aligned}
&= S_y^2 - 2B_1 R S_x S_y + B_1^2 S_x^2 \\
&= S_y^2 - 2\frac{R^2 S_y}{S_x} S_x S_y + \frac{R^2 S_y^2}{S_x^2} S_x^2 \\
&= S_y^2 (1 - R^2).
\end{aligned}
$$

4.32 (a) Since both $\bar{x}$ and $\bar{y}$ are unbiased for their respective population means, $E[\hat{\bar{y}}_{\text{diff}}] = \bar{y}_{\mathcal{U}} + (\bar{x}_{\mathcal{U}} - \bar{x}_{\mathcal{U}}) = \bar{y}_{\mathcal{U}}$.

(b) According to Raj (1968),

$$
V(\hat{\bar{y}}_{\text{diff}}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \left(S_y^2 + S_x^2 - 2RS_x S_y\right)
$$

Or, defining the residuals $e_i = y_i - x_i$,

$$
V(\hat{\bar{y}}_{\text{diff}}) = \left(1 - \frac{n}{N}\right) \frac{S_e^2}{n}
$$

The estimated total difference is $\hat{t}_y - \hat{t}_x = N(\bar{y} - \bar{x})$; the estimated audited value for accounts receivable is

$$
\hat{t}_{y\text{diff}} = t_x + (\hat{t}_y - \hat{t}_x).
$$

The residuals from this model are $e_i = y_i - x_i$. The variance of $\hat{t}_{y\text{diff}}$ is

$$
V(\hat{t}_{y\text{diff}}) = V[t_x + (\hat{t}_y - \hat{t}_x)] = V(\hat{t}_e),
$$

where $\hat{t}_e = (N/n) \sum_{i \in \mathcal{S}} e_i$. If the variability in the residuals e_i is smaller than the variability among the y_i, then difference estimation will increase precision.

Difference estimation works best if the population and sample have a large fraction of nonzero differences that are roughly equally divided between overstatements and understatements, and if the sample is large enough so that the sampling distribution of $(\bar{y} - \bar{x})$ is approximately normal. In auditing, it is possible that most of the audited values in the sample are all the same as the corresponding book value. In such a situation, the design-based variance estimator is unstable and a model-based approach may be preferred.

4.35 From linear models theory, if $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E[\boldsymbol{\varepsilon}] = 0$ and $\text{Cov}[\varepsilon] = \sigma^2 \mathbf{A}$, 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, 2020). In our case,

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

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

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

4.38 (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 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E[\boldsymbol{\varepsilon}] = 0$ 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 The case fatality rate is calculated by dividing the estimated number of fatalities by the estimated number of cases. Both numerator and denominator are subject to sampling and nonsampling error. By contrast, the denominator for the population mortality rate is the population of a region, which is usually known with much greater accuracy.

4.44

```
/* Part (a) */
proc surveymeans  data = vius plots = none mean sum clm clsum;
   weight tabtrucks;
   strata stratum;
   var miles_annl;
   domain business;

/* Parts (b,c) */
proc surveymeans  data = vius  plots = none mean clm ;
   weight tabtrucks;
   strata stratum;
   var mpg;
   ratio miles_annl/miles_life;
   domain transmssn;
run;
```

business	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum
1	56452	1246.317720	54009.5907	58895.1519	72272793289	1608230919
2	23306	790.981413	21755.6067	24856.2511	20024589014	1213307392
3	10768	415.312729	9954.2886	11582.3131	24119946651	1354386330
4	19210	1126.305405	17002.3551	21417.4684	3411543277	360265917
5	15081	830.112607	13454.3175	16708.3561	10244675655	942933274
6	16714	381.173223	15966.8537	17461.0515	75906142636	2821651145
7	19650	1018.600046	17653.3448	21646.2535	15384530602	1399406209
8	23052	1184.715817	20729.7496	25373.8316	16963450921	1348917090
9	17948	561.147469	16848.3582	19048.0543	27470445448	1422261638
10	14927	1396.653144	12189.9160	17664.7915	5622014452	923917751
11	14410	726.842687	12985.5076	15834.7285	10709275945	901989658
12	9537	1267.485898	7052.3180	12020.8583	1784083855	353650310
13	20461	1300.857231	17911.2857	23010.6416	5816313888	677802928
14	16818	600.239019	15641.1955	17994.1304	35776203775	2201296141

transmssn	Variable	Mean	Std Error of Mean	95% CL for Mean	
1	mpg	16.665300	0.043656	16.580	16.751
2	mpg	16.022122	0.097739	15.831	16.214
3	mpg	14.846222	1.012044	12.863	16.830
4	mpg	16.732086	1.599588	13.597	19.867

Numerator	Denominator	Ratio	Std Error	95% CL for Ratio	
miles_annl	miles_life	0.124415	0.000932	0.12258719	0.12624229

(c) 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.

5.2 (a) This is a cluster sample because there are two sizes of sampling units: the clinics are the psus and parents who attended the clinics are ssus. It is a one-stage sample because questionnaires were distributed to all parents visiting the selected clinics.

(b) This is not a probability sample. The clinics were chosen conveniently. It is certainly not a representative sample of households with children because, even if the clinics were a representative sample, many parents do not take their children to a clinic.

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

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

From the data file,

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

and

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

Then, using (5.20),

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

The estimated variance of the residuals is

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

using (5.17),

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

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 Here is code and output from R:

```
data(spanish)
spanish$popsize <- rep(72,nrow(spanish))
dspan <- svydesign(id=~class,fpc=~popsize,data=spanish)
degf(dspan)
## [1] 9
mscore<-svymean(~trip+score,dspan)
mscore
```

```
##           mean     SE
## trip   0.32143 0.0733
## score 66.79592 2.7091
confint(mscore,df=degf(dspan))
##            2.5 %     97.5 %
## trip   0.1555907  0.4872665
## score 60.6675152 72.9243216
svytotal(~trip+score,dspan)
##         total      SE
## trip    453.6  111.82
## score 94262.4 6775.43
```

The following code for SAS software gives the same results. Since this is a one-stage cluster sample, there is no contribution to variance from subsampling. The fpc computed by the SURVEYMEANS procedure gives the correct without-replacement variance.

```
data spanish;
   set datalib.spanish;
   sampwt = 72/10;
   /* weight = N/n = 72/10 since one-stage cluster sample*/

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

Note that the sum of the weights is 1411.2.

5.6 (a) The R code and output below was used to calculate summary statistics and the ANOVA table.

```
worms<-data.frame(
     case= rep(c(1:12),each=3),
     can = rep(c(1:3),times=12),
     worms = c(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)
     )
worms$Mi<-rep(24,nrow(worms))
worms$N<-rep(580,nrow(worms))
# worms # check data entry--using rep can be tricky!
dworms<-svydesign(id=~case+can,fpc=~N+Mi,data=worms)
mworms<-svymean(~worms,dworms)
mworms
##
##        mean     SE
## worms 3.6389 0.6102
confint(mworms,df=degf(dworms)))
```

```
##          2.5 %   97.5 %
## worms 2.295864 4.981913
```

From the output, we see that

$$\hat{\bar{y}}_{\text{unb}} = 3.6389$$

with standard error 0.61.

If desired, you can also look at the contributions from each term using the formula.

```
wormstat<-data.frame(meanworms=tapply(worms$worms,worms$case,mean),
           varworms=tapply(worms$worms,worms$case,var))
wormstat
   meanworms  varworms
1   4.333333  9.333333
2   3.333333  1.333333
3   1.000000  1.000000
4   5.000000  3.000000
5   7.000000  7.000000
6   3.333333 12.333333
7   3.666667  5.333333
8   1.666667  2.333333
9   5.000000  4.000000
10  2.666667  2.333333
11  6.666667  6.333333
12  0.000000  0.000000

varformula_t1<- (1-12/580)*var(3*wormstat$meanworms)/(12*3^2)
varformula_t2<- (1/580)*(1-3/24)*mean(wormstat$varworms)/3
varformula_t1
## [1] 0.3700579
varformula_t2
## [1] 0.0022769
anova(lm(worms~factor(case),data=worms))
##
Analysis of Variance Table

Response: worms
             Df Sum Sq Mean Sq F value  Pr(>F)
factor(case) 11 149.64 13.6035  3.0045 0.01172 *
Residuals    24 108.67  4.5278
---
```

$$\text{SE}\left[\hat{\bar{y}}_{\text{unb}}\right] = \sqrt{\left(1 - \frac{12}{580}\right)\frac{13.604}{(12)(3)^2} + \frac{1}{580}\left(1 - \frac{3}{24}\right)\frac{4.528}{3}}$$

$$= \sqrt{0.3701 + 0.0023}$$
$$= 0.61.$$

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

5.7 Here are statistics from R.

```
cases<-c(146, 180, 251, 152, 72, 181, 171, 361, 73, 186, 99, 101, 52, 121, 199, 179, 98,
    63, 126, 87, 62,
226, 129, 57, 46, 86, 43, 85, 165, 12, 23, 87, 43, 59)
Mi<-c(52,19,37,39,8,14)
mi<-c(10,4,7,8,2,3)
glob<-data.frame(city=rep(1:6,mi),Mi=rep(Mi,mi),mi=rep(mi,mi),cases)
glob$smkt<-1:length(cases) #supermarket id
glob$wt<-(45/6)*glob$Mi/glob$mi
glob$wt
##  [1] 39.00000 39.00000 39.00000 39.00000 39.00000 39.00000 39.00000 39.00000
##  [9] 39.00000 39.00000 35.62500 35.62500 35.62500 35.62500 39.64286 39.64286
## [17] 39.64286 39.64286 39.64286 39.64286 39.64286 36.56250 36.56250 36.56250
## [25] 36.56250 36.56250 36.56250 36.56250 36.56250 30.00000 30.00000 35.00000
## [33] 35.00000 35.00000
# glob
# tapply(glob$cases,glob$city,mean)
# tapply(glob$cases,glob$city,var)
# with-replacement variance
dglobwr<-svydesign(id=~city,weights=~wt,data=glob)
svymean(~cases,dglobwr)
##         mean     SE
## cases 120.69 21.194
svytotal(~cases,dglobwr)
##       total    SE
## cases 152972 60800
# without-replacement variance
dglobwor<-svydesign(id=~city+smkt,fpc=~rep(45,length(cases))+Mi,data=glob)
svymean(~cases,dglobwor)
##         mean     SE
## cases 120.69 20.049
svytotal(~cases,dglobwor)
##       total    SE
## cases 152972 56781
```

If you prefer the formulas for the without-replacement variance, we can use (5.21) and (5.27) for calculating the total number of cases sold, and (5.30) and (5.32) for calculating the average number of cases sold per supermarket. Summary quantities are given in the following table.

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 (5.21),

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

Using (5.27),

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

From (5.30) and (5.32),

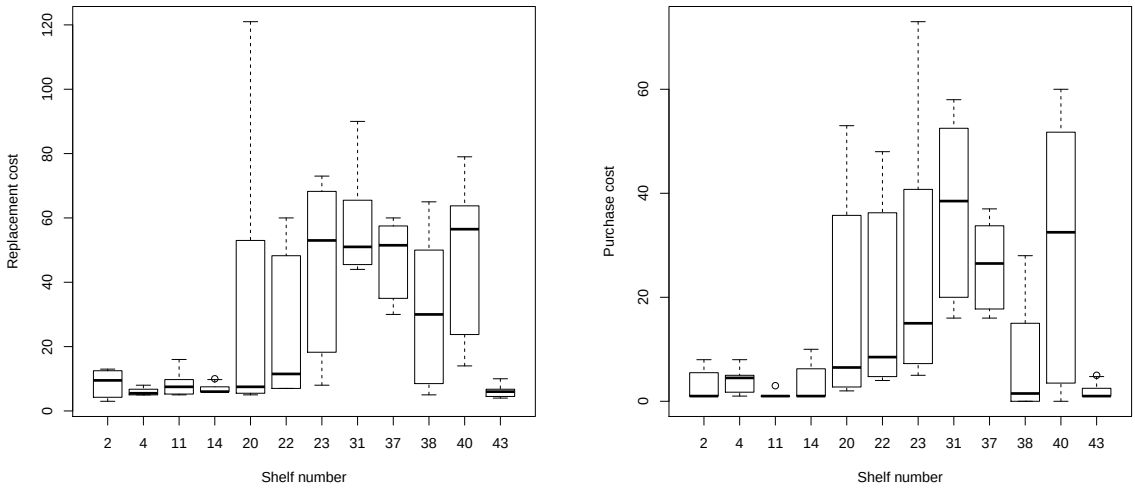
$$\hat{\bar{y}}_r = \frac{20{,}396}{169} = 120.7$$

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, 5.9 The R code below computes estimates and draws plots for both problems

```
data(books)
dbooks<-svydesign(id=~shelf+booknumber,fpc=~rep(44,60)+Mi,nest=TRUE,data=books)
par(las=1)
svyboxplot(replace~factor(shelf),dbooks,xlab="Shelf number",ylab="Replacement cost")
svyboxplot(purchase~factor(shelf),dbooks,xlab="Shelf number",ylab="Purchase cost")
svytotal(~replace+purchase,dbooks)
##          total     SE
## replace  32638 5733.5
## purchase 17891 3977.9
svymean(~replace+purchase,dbooks)
##            mean     SE
## replace  23.611 5.4759
## purchase 12.943 3.6994
```



It appears that the means and variances differ quite a bit for different shelves, for each response variable.

If you want to see the formulas, 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

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

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

$$\mathrm{SE}\,(\hat{t}_{\mathrm{unb}})/\hat{t}_{\mathrm{unb}} = \frac{5733.52}{32637.73} = 0.176.$$

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

```
                          Std Error
Variable           Mean        of Mean      95% CL for Mean
replace       23.610610       6.344128   9.64727746 37.5739427


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.

(c) To find the average replacement cost per book with the information given, we use the ratio estimate in (5.30):

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

In the formula for variance in (5.32),

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

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

$$\mathrm{SE}\,(\hat{\bar{y}}_r) = \frac{1}{31.417}\sqrt{\left(1-\frac{12}{44}\right)\frac{476718.4455}{12}+705.4}$$

$$= \frac{1}{31.417}\sqrt{28892.0 + 705.4} = 5.476.$$

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 See 5.8

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 (5.13) to estimate R_a^2: we have

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

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

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

so

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

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),

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

(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 5.4.

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

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, (5.33) yields

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

5.13 (a) Cluster sampling is needed for this application because the household is the sampling unit. 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. From the equation following (5.13),

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

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 Table SM5.5.1.

Table 5.1: Calculations for part (b) of Exercise 5.15.

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

Table 5.2: Calculations for Exercise 5.15(d)

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	

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 Table SM5.5.2.

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

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 (5.30) and (5.32),

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

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

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 $\hat{V}(\hat{p}) = .569(1-.569)/280 = .000876$.

5.16 (a) Table SM5.3 gives summary quantities; the column $\bar{y}_i$ gives the estimated proportion of children who had previously had measles in each school.

(b) $\hat{\bar{y}}_r = \dfrac{863.41}{1961} = 0.4403$

$$\begin{aligned}\hat{V}(\hat{\bar{y}}_r) &= \frac{1}{(196.1)^2}\left[\left(1-\frac{10}{46}\right)\frac{1890.15}{10} + \frac{2888.16}{(10)(46)}\right] \\ &= \frac{1}{(196.1)^2}[147.92 + 6.28] = 0.004.\end{aligned}$$

Table 5.3: Calculations for Exercise 5.16

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	

An approximate 95% CI is

$$0.4403 \pm 1.96\sqrt{0.004} = [0.316, 0.564]$$

5.18 Here is R code and output for analyzing *reading*:

```
data(schools)
# calculate with-replacement variance; no fpc argument
# include psu variable in id; include weights
dschools<-svydesign(id=~schoolid,weights=~finalwt,data=schools)
readmean<-svymean(~reading,dschools)
readmean
##           mean     SE
## reading 30.604 1.5533
confint(readmean,df=degf(dschools))
##            2.5 %   97.5 %
## reading 27.09028 34.11797
# estimate proportion and total number of students with mathlevel=2
rlevel2<-svymean(~factor(readlevel),dschools)
rlevel2
##                          mean     SE
## factor(readlevel)1 0.52171 0.0657
## factor(readlevel)2 0.47829 0.0657
confint(rlevel2,df=degf(dschools))
##                           2.5 %    97.5 %
## factor(readlevel)1 0.3730626 0.6703590
## factor(readlevel)2 0.3296410 0.6269374
```

```
r2total<-svytotal(~factor(readlevel),dschools)
r2total
##                     total     SE
## factor(readlevel)1 9011.2 1954.2
## factor(readlevel)2 8261.2 1032.9
confint(r2total,df=degf(dschools))
##                         2.5 %   97.5 %
## factor(readlevel)1 4590.461 13432.04
## factor(readlevel)2 5924.756 10597.74
# without replacement
schools$studentid<-1:(nrow(schools))
dschoolwor<-svydesign(id=~schoolid+studentid,fpc=~rep(75,nrow(schools))+Mi,
            data=schools)
readmeanwor<-svymean(~reading,dschoolwor)
readmeanwor
##           mean     SE
## reading 30.604 1.4612
confint(readmeanwor,df=degf(dschoolwor))
##            2.5 %   97.5 %
## reading 27.29872 33.90953

# estimate proportion and total number of students with mathlevel=2
svymean(~factor(readlevel),dschoolwor)
                      mean     SE
factor(readlevel)1 0.52171 0.0624
factor(readlevel)2 0.47829 0.0624
> svytotal(~factor(readlevel),dschoolwor)
                     total      SE
factor(readlevel)1 9011.2 1832.11
factor(readlevel)2 8261.2  985.56
# Problem 5.20
library(nlme)
readmixed <- lme(fixed=reading~1,random=~1|factor(schoolid),data=schools)
summary(readmixed)
## Linear mixed-effects model fit by REML
##   Data: schools
##       AIC     BIC   logLik
##   1404.03 1413.91 -699.015

## Random effects:
##  Formula: ~1 | factor(schoolid)
##         (Intercept) Residual
## StdDev:    4.373528  7.65006
```

```
## Fixed effects:  reading ~ 1
##               Value Std.Error  DF  t-value p-value
## (Intercept) 31.565  1.485056 190 21.25509       0

## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.5151010 -0.6911411  0.1162636  0.8693088  1.6368259

## Number of Observations: 200
## Number of Groups: 10
# extract the variance components
VarCorr(readmixed)
## factor(schoolid) = pdLogChol(1)
##             Variance StdDev
## (Intercept) 19.12775 4.373528
## Residual    58.52342 7.650060
```

(a) The mean reading score is 30.604. The with-replacement SE is 1.5533.

(b) The without-replacement SE is 1.4612, just a little bit smaller.

5.19 From the output in the previous exercise, the estimated proportion is 0.47829 with SE 0.0657 and 95% CI [0.330, 0.627]. The CI is quite wide!

5.20 From the R code for Exercise 5.18, we have $\hat{\mu} = 31.565$ with SE 1.485.

Here is analogous code for the MIXED procedure in SAS software:

```
proc mixed data=schools;
   class schoolid;
   model reading = / solution cl;
   random schoolid;
run;
```

5.21 With the different costs, the last two columns of Table 5.8 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.

5.22 (a) Here is the sample ANOVA table from the GLM procedure of SAS software:

Source	DF	Sum of 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 (5.36), 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 $n = 300/4 = 75$.

5.23 Answers will vary.

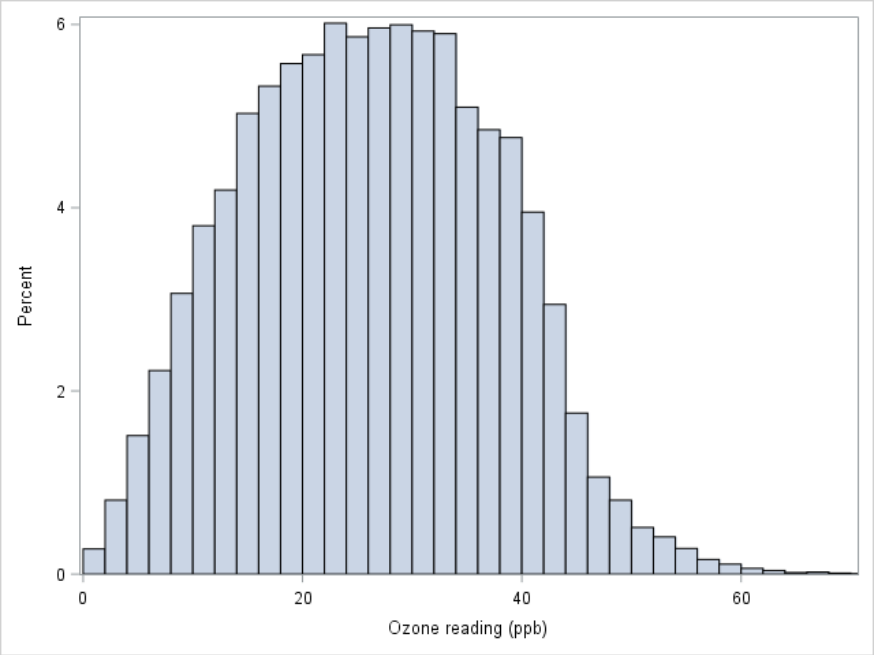
5.24 Students using the .csv file must change the missing value code from -9 before doing calculations with the data.

(a) Figure SM5.1 shows a histogram for the population values.

The population has mean 25.782, standard deviation 11.3730526, and median 26. These are calculated ignoring the small amount of missing data.

(b, c) Answers will vary, depending on which hour is chosen for the systematic sample. Table SM5.4 gives the summary statistics for each of the 24 possible systematic samples.

Figure 5.1: Histogram of ozone readings for site in Monterey County



5.25 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

$$\begin{aligned}\text{ICC} &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_U)(y_{ik}-\bar{y}_U)}{(NM-1)(M-1)S^2} \\ &= \frac{M(\text{SSB})-\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2}{(NM-1)(M-1)S^2} \\ &= \frac{M(\text{SSB})-\text{SSTO}}{(M-1)\text{SSTO}} \\ &= \frac{M(\text{SSTO}-\text{SSW})-\text{SSTO}}{(M-1)\text{SSTO}} \\ &= 1-\frac{M}{M-1}\,\frac{\text{SSW}}{\text{SSTO}}.\end{aligned}$$

Table 5.4: Summary statistics for the 24 possible systematic samples.

Hour	Mean	Std Dev	Median
hr0	19.910	9.639	19
hr1	19.486	9.735	18
hr2	19.166	9.862	18
hr3	18.740	10.027	17
hr4	18.186	10.045	17
hr5	17.629	9.914	16
hr6	17.488	9.739	16
hr7	19.081	9.315	18
hr8	22.381	9.205	22
hr9	26.734	9.346	27
hr10	31.092	9.186	31
hr11	34.347	8.789	34
hr12	35.729	8.508	35
hr13	36.108	8.391	36
hr14	36.013	8.402	36
hr15	35.492	8.640	35
hr16	34.218	8.615	34
hr17	31.853	8.915	32
hr18	29.058	8.741	29
hr19	26.999	8.419	27
hr20	24.701	8.508	24
hr21	22.949	8.751	22
hr22	21.501	9.048	21
hr23	20.596	9.540	20

5.26 From (5.10),

$$\text{ICC} = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}.$$

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 (5.10),

$$\begin{aligned} \text{ICC} &= 1 - \frac{M}{M-1}\frac{\text{SSTO}-(N-1)\text{MSB}}{\text{SSTO}} \\ &= -\frac{1}{M-1} + \frac{M(N-1)\text{MSB}}{(M-1)(NM-1)S^2} \end{aligned}$$

and

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

5.27 (a) From (5.27), 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 (5.9) 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 (5.13). 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.

5.28 (a) If $M_i = M$ for all i, and $m_i = m$ for all i, then

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} M\bar{y}_i}{nM} = \frac{1}{NM}\,\hat{t}_{\text{unb}} = \hat{\bar{y}}_{\text{unb}}.$$

(b) In the following, let $\hat{\bar{y}} = \hat{\bar{y}}_r = \hat{\bar{y}}_{\text{unb}}$

Source	df	SS
Between clusters	$n-1$	$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(\bar{y}_i - \hat{\bar{y}})^2$
Within clusters	$n(m-1)$	$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2$
Total	$nm-1$	$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \hat{\bar{y}})^2$

(c) Let

$$Z_i = \begin{cases} 1 & \text{if psu } i \text{ in sample} \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned}
\widehat{\text{SSW}} &= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2 \\
E[\widehat{\text{SSW}}] &= E\Big[\sum_{i=1}^{N} Z_i \sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2\Big] \\
&= E\Big\{\sum_{i=1}^{N} Z_i E\Big[\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2 \mid \mathbf{Z}\Big]\Big\} \\
&= E\Big\{\sum_{i=1}^{N} Z_i (m-1) E[s_i^2 \mid \mathbf{Z}]\Big\} \\
&= E\Big\{\sum_{i=1}^{N} Z_i (m-1) S_i^2\Big\} \\
&= (m-1)\frac{n}{N}\sum_{i=1}^{N} S_i^2
\end{aligned}$$

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,

$$\hat{\bar{y}}_{\text{unb}} = \frac{1}{n}\sum_{i\in\mathcal{S}} \bar{y}_i = \frac{1}{n}\sum_{i=1}^{N} Z_i \bar{y}_i$$

$$\widehat{\text{SSB}} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} (\bar{y}_i - \hat{\bar{y}}_{\text{unb}})^2 = m\sum_{i\in\mathcal{S}} (\bar{y}_i - \hat{\bar{y}}_{\text{unb}})^2$$

$$\begin{aligned}
E[\widehat{\text{SSB}}] &= mE\Big[\sum_{i\in\mathcal{S}} (\bar{y}_i^2 - 2\bar{y}_i\hat{\bar{y}}_{\text{unb}} + \hat{\bar{y}}_{\text{unb}}^2)\Big] \\
&= mE\Big[\sum_{i\in\mathcal{S}} \bar{y}_i^2 - n\hat{\bar{y}}_{\text{unb}}^2\Big] \\
&= mE\Big(E\Big[\sum_{i=1}^{N} Z_i\bar{y}_i^2 \mid Z_1,\ldots,Z_n\Big]\Big) - mnE[\hat{\bar{y}}_{\text{unb}}^2] \\
&= mE\Big[\sum_{i=1}^{N} Z_i\{V(\bar{y}_i \mid \mathbf{Z}) + \bar{y}_{iU}^2\}\Big] - mn[V(\hat{\bar{y}}_{\text{unb}}) + \bar{y}_U^2] \\
&= mE\Big[\sum_{i=1}^{N} Z_i\Big\{\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} + \bar{y}_{iU}^2\Big\}\Big] \\
&\quad -mn\Big[\frac{1}{M^2}\Big(1-\frac{n}{N}\Big)\frac{S_t^2}{n} + \frac{1}{nN}\sum_{i=1}^{N}\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} + \bar{y}_U^2\Big] \\
&= \frac{mn}{N}\sum_{i=1}^{N}\Big[\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} + \bar{y}_{iU}^2\Big] - \frac{m}{M^2}\Big(1-\frac{n}{N}\Big)S_t^2 \\
&\quad -\frac{m}{N}\sum_{i=1}^{N}\Big(1-\frac{m}{M}\Big)\frac{S_i^2}{m} - mn\bar{y}_U^2 \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + mn\Big[\frac{1}{N}\sum_{i=1}^{N}\bar{y}_{iU}^2 - \bar{y}_U^2\Big] \\
&\quad -\frac{m}{M}\Big(1-\frac{n}{N}\Big)(\text{MSB}) \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + \frac{mn}{N}\sum_{i=1}^{N}(\bar{y}_{iU} - \bar{y}_U)^2 \\
&\quad -\frac{m}{M}\Big(1-\frac{n}{N}\Big)\frac{1}{N-1}\text{SSB} \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + \Big[\frac{mn}{NM} - \frac{m}{M(N-1)}\Big(1-\frac{n}{N}\Big)\Big]\text{SSB} \\
&= \frac{m(n-1)}{N}\Big(1-\frac{m}{M}\Big)\sum_{i=1}^{N}\frac{S_i^2}{m} + \frac{(N-1)mn - m(N-n)}{NM(N-1)}\text{SSB}
\end{aligned}$$

$$= \quad (n-1)\Big(1-\frac{m}{M}\Big)\text{MSW} + (n-1)\frac{m}{M}\text{MSB}.$$

Thus

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

(d) From (c),

$$E[\widehat{\text{MSB}}] = \frac{M}{m}\Big(1-\frac{m}{M}\Big)\text{MSW} + \frac{M}{m}\frac{m}{M}\text{MSB} - \Big(\frac{M}{m}-1\Big)\,\text{MSW} = \text{MSB}.$$

(e)

$$\begin{aligned} s_t^2 &= \frac{1}{n-1}\sum_{i\in\mathcal{S}}\Big(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\Big)^2 \\ &= \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M\bar{y}_i - M\hat{\bar{y}})^2 \\ &= \frac{1}{n-1}\frac{M^2}{m}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(\bar{y}_i - \hat{\bar{y}})^2 \\ &= \frac{M^2}{m}\text{msb} \end{aligned}$$

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 (5.27),

$$\begin{aligned} \hat{V}(\hat{\bar{y}}) &= \frac{1}{(NM)^2}\Big[N^2\Big(1-\frac{n}{N}\Big)\frac{s_t^2}{n} + \frac{N}{n}\Big(1-\frac{m}{M}\Big)\frac{M^2}{m}\sum_{i\in\mathcal{S}} s_i^2\Big] \\ &= \Big(1-\frac{n}{N}\Big)\frac{\text{msb}}{nm} + \frac{1}{N}\Big(1-\frac{m}{M}\Big)\frac{\text{msw}}{m}. \end{aligned}$$

5.29 (a) From Exercise 5.28, $\text{msto} = \frac{1}{nm-1}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij}-\hat{\bar{y}})^2$. Now,

$$\begin{aligned} & E\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(y_{ij}-\hat{\bar{y}})^2\right] \\ &= \quad E[\widehat{\text{SSW}}] + E[\widehat{\text{SSB}}] \\ &= \quad (m-1)\frac{n}{N}\sum_{i=1}^{N} S_i^2 + (n-1)\Big(1-\frac{m}{M}\Big)\text{MSW} + (n-1)\frac{m}{M}\text{MSB} \end{aligned}$$

$$
\begin{aligned}
&= \left[(m-1)n + (n-1)\left(1-\frac{m}{M}\right)\right]\text{MSW} + (n-1)\frac{m}{M}\text{MSB} \\
&= \left[nm - 1 - \frac{m}{M}(n-1)\right]\text{MSW} + (n-1)\frac{m}{M}\text{MSB} \\
&= \frac{1}{N(M-1)}\left[nm - 1 - \frac{m}{M}(n-1)\right]\text{SSW} + \frac{(n-1)}{N-1}\frac{m}{M}\text{SSB} \\
&= \frac{nm-1}{NM-1}\left[\frac{NM-1}{N(M-1)}\text{SSW} + \frac{(n-1)m}{(N-1)M}\frac{NM-1}{nm-1}\text{SSB}\right] - \frac{m(n-1)}{NM(M-1)}\text{SSW} \\
&= \frac{nm-1}{NM-1}\left[\left\{1 + \frac{N-1}{N(M-1)}\right\}\text{SSW}\right. \\
&\qquad \left. + \left\{1 + \frac{nm(M-1) - (m-1)NM - M + m}{(N-1)M(nm-1)}\right\}\text{SSB}\right] \\
&\quad -\frac{m(n-1)}{NM(M-1)}\text{SSW} \\
&= \frac{nm-1}{NM-1}\text{SSTO} + \frac{nm-1}{NM-1}\frac{nm(M-1) - (m-1)NM - M + m}{(N-1)M(nm-1)}\text{SSB} \\
&\quad + \left[\frac{nm-1}{NM-1}\frac{N-1}{N(M-1)} - \frac{m(n-1)}{NM(M-1)}\right]\text{SSW} \\
&= \frac{nm-1}{NM-1}\left[\text{SSTO} + O\left(\frac{1}{n}\right)\text{SSB} + O\left(\frac{1}{n}\right)\text{SSW}\right].
\end{aligned}
$$

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 5.28, $E[\text{msw}] = \text{MSW}$ and

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

Consequently,

$$
\begin{aligned}
E[\hat{S}^2] &= \frac{M(N-1)}{m(NM-1)}E[\text{msb}] + \frac{(m-1)NM + M - m}{m(NM-1)}E[\text{msw}] \\
&= \frac{M(N-1)}{m(NM-1)}\left[\frac{m}{M}\text{MSB} + \left(1-\frac{m}{M}\right)\text{MSW}\right] + \frac{(m-1)NM + M - m}{m(NM-1)}\text{MSW} \\
&= \frac{1}{NM-1}\text{SSB} + \frac{1}{NM-1}\left[(N-1)(M-m) + (m-1)NM + M - m\right]\text{MSW} \\
&= \frac{1}{NM-1}\text{SSB} + \frac{N(M-1)}{NM-1}\text{MSW} \\
&= S^2.
\end{aligned}
$$

5.30 The cost constraint implies that $n = C/(c_1 + c_2 m)$; substituting into (5.34), we have:

$$g(m) = V(\hat{\bar{y}}_{\text{unb}}) = \frac{(c_1 + c_2 m)\text{MSB}}{CM} - \frac{\text{MSB}}{NM} + \left(1 - \frac{m}{M}\right)\frac{(c_1 + c_2 m)\text{MSW}}{C_m}$$
$$\frac{dg}{dm} = \frac{c_2 \text{MSB}}{CM} - \frac{c_1 \text{MSW}}{Cm^2} - \frac{c_2 \text{MSW}}{CM}.$$

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 5.27,

$$m = \sqrt{\frac{c_1 M S^2(1 - R_a^2)}{c_2 S^2[N(M-1)R_a^2/(N-1) + 1 - (1 - R_a^2)]}} = \sqrt{\frac{c_1 M(N-1)(1 - R_a^2)}{c_2(NM - 1)R_a^2}}.$$

5.31 Using (5.28) and (5.24),

$$\begin{aligned} E\left[\hat{V}_{\text{psu}}\left(\hat{t}_{\text{unb}}\right)\right] &= N^2\left(1 - \frac{n}{N}\right)\frac{S_t^2}{n} + N^2\left(1 - \frac{n}{N}\right)\frac{1}{N}\sum_{i=1}^{N} V\left(\hat{t}_i\right) \\ &= V\left(t_{\text{unb}}\right) - \frac{N}{n}\sum_{i=1}^{N} V\left(\hat{t}_i\right) + N\left(1 - \frac{n}{N}\right)\sum_{i=1}^{N} V\left(\hat{t}_i\right) \\ &= V\left(t_{\text{unb}}\right) - \sum_{i=1}^{N} V\left(\hat{t}_i\right). \end{aligned}$$

Similarly,

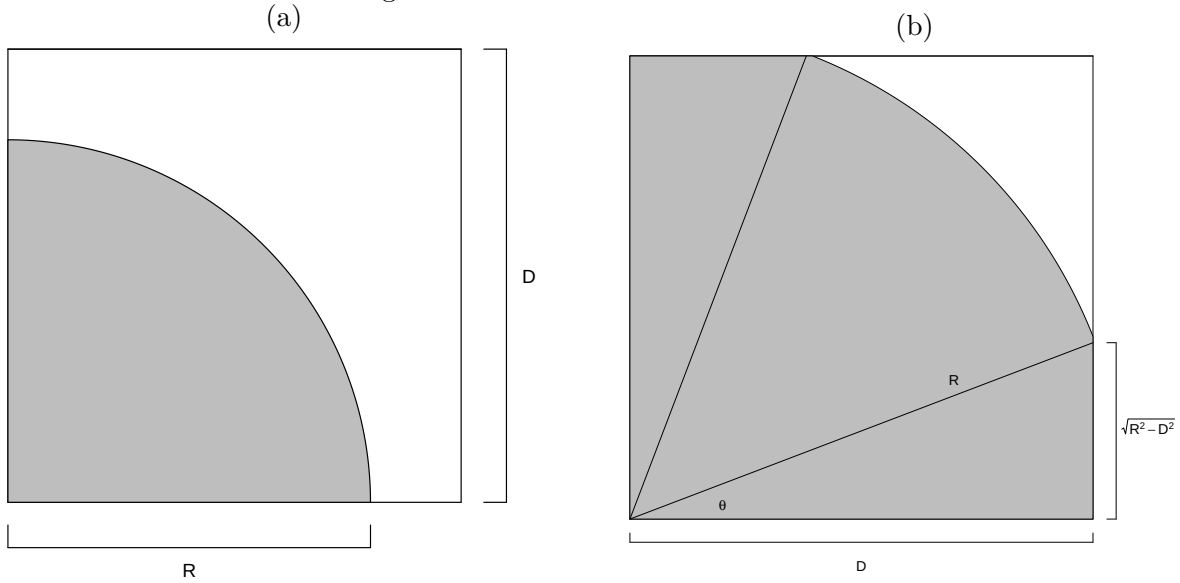
$$\begin{aligned} E\left[\hat{V}_{\text{WR}}\left(\hat{t}_{\text{unb}}\right)\right] &= N^2\frac{S_t^2}{n} + \frac{N^2}{n}\frac{1}{N}\sum_{i=1}^{N} V\left(\hat{t}_i\right) \\ &= V\left(t_{\text{unb}}\right) + NS_t^2 - \frac{N}{n}\sum_{i=1}^{N} V\left(\hat{t}_i\right) + \frac{N}{n}\sum_{i=1}^{N} V\left(\hat{t}_i\right) \\ &= V\left(t_{\text{unb}}\right) + NS_t^2. \end{aligned}$$

5.32 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: $R < D$. See Figure SM5.2(a).

Figure 5.2: Pictures for Exercise 5.32.



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: $D \leq R \leq \sqrt{2}D$. See Figure SM5.2(b).

The probability is again

$$\frac{\text{area of shaded part}}{\text{area of square}}.$$

The area of the shaded part is

$$2\left(\frac{1}{2}D\sqrt{R^2 - D^2}\right) + \left(\frac{\pi/2 - 2\theta}{\pi/2}\right)\left(\frac{\pi R^2}{4}\right) = D\sqrt{R^2 - D^2} + \left(1 - \frac{4\theta}{\pi}\right)\left(\frac{\pi R^2}{4}\right),$$

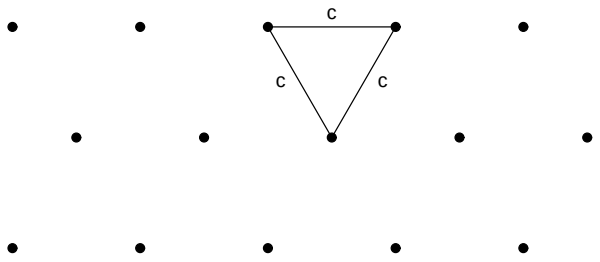
where $\cos(\theta) = D/R$.

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.



5.33 (a) We have msb = 0.56392 and msw= $\hat{\sigma}^2 = 0.18504$, so $\hat{\sigma}_A^2 = (0.56392 - 0.18504)/4 = 0.09472$. Thus, $\hat{\rho} = 0.09472/(0.09472 + 0.18504) = 0.3386$. It is almost the same as the estimated ICC. Note that $\hat{\rho}$ can also be estimated using the MIXED procedure in SAS software or *lmer* in R—in this case, the method of moments and REML estimates are the same because every cluster has the same size.

(b) For Population A, msb − msw < 0. Since we can't have a negative variance, we take $\hat{\sigma}_A = 0$ and $\hat{\rho} = 0$. But the estimated ICC was negative.

For Population B, we have $\hat{\sigma}_A^2 = 110.44$ and $\hat{\sigma}^2 = 1.6667$, so $\hat{\rho} = 110.44/(110.44 + 1.667) = 0.9851$.

5.34 Because the A_is and the ε_{ij}s are independent,

$$\begin{aligned} V_{\mathrm{M1}}\Big[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}Y_{ij}\Big] &= V_{\mathrm{M1}}\Big[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}(\mu + A_i + \varepsilon_{ij})\Big] \\ &= V_{\mathrm{M1}}\Big[\sum_{i\in\mathcal{S}} A_i\Big(\sum_{j\in\mathcal{S}_i} b_{ij}\Big)\Big] + V_{\mathrm{M1}}\Big[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}\varepsilon_{ij}\Big] \\ &= \sum_{i\in\mathcal{S}}\Big(\sum_{j\in\mathcal{S}_i} b_{ij}\Big)^2\sigma_A^2 + \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}^2\sigma^2. \end{aligned}$$

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), and noting that under

model M1 all A_i and ε_{ij} are independent,

$$\begin{aligned}
V_{\mathrm{M1}}[\hat{T}-T] &= \sum_{i=1}^{N}\Big(\sum_{j=1}^{M_i} c_{ij}\Big)^2\sigma_A^2 + \sum_{i=1}^{N}\sum_{j=1}^{M_i} c_{ij}^2\sigma^2 \\
&= \sum_{i\in\mathcal{S}}\Big[\Big(\sum_{j\in\mathcal{S}_i} c_{ij} + \sum_{j\notin\mathcal{S}_i} c_{ij}\Big)^2\Big]\sigma_A^2 + \sum_{i\notin\mathcal{S}}\Big(\sum_{j=1}^{M_i} c_{ij}\Big)^2\sigma_A^2 \\
&\quad + \sum_{i\in\mathcal{S}}\Big(\sum_{j\in\mathcal{S}_i} c_{ij}^2 + \sum_{j\notin\mathcal{S}_i} c_{ij}^2\Big)\sigma^2 + \sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i} c_{ij}^2\sigma^2 \\
&= \sum_{i\in\mathcal{S}}\Big[\Big\{\sum_{j\in\mathcal{S}_i}(b_{ij}-1) - (M_i-m_i)\Big\}^2\Big]\sigma_A^2 + \sum_{i\notin\mathcal{S}} M_i^2\sigma_A^2 \\
&\quad + \sum_{i\in\mathcal{S}}\Big[\sum_{j\in\mathcal{S}_i}(b_{ij}-1)^2 + (M_i-m_i)\Big]\sigma^2 + \sum_{i\notin\mathcal{S}} M_i\sigma^2 \\
&= \sum_{i\in\mathcal{S}}\Big[\sum_{j\in\mathcal{S}_i} b_{ij} - M_i\Big]^2\sigma_A^2 + \sum_{i\notin\mathcal{S}} M_i^2\sigma_A^2 \\
&\quad + \sum_{i\in\mathcal{S}}\Big[\sum_{j\in\mathcal{S}_i}(b_{ij}^2 - 2b_{ij}) + M_i\Big]\sigma^2 + \sum_{i\notin\mathcal{S}} M_i\sigma^2
\end{aligned}$$

5.35 Let $\mathbf{Y}_{\mathcal{S}} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \mathbf{Y}_n^T)^T$ denote the vector of random variables observed in the sample. For convenience, we label the sampled psus from 1 to n, and we label the sampled ssus within sampled psus from $1, \ldots, m_i$ so that the vector of observations from psu i is $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i,m_i})^T$.

Then, under model M1,

$$\mathbf{Y}_i \sim N\left(\mathbf{1}_{M_0}\mu, \mathit{blockdiag}\left[\sigma_A^2\mathbf{1}_{M_i}\mathbf{1}_{M_i}^T + \sigma^2\mathbf{I}_{M_i}\right]\right),$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{1}_k$ is the vector of length k with all entries equal to one.

Let $\mathbf{Y}_{\mathcal{S}}$ denote the set $\{Y_{ij} : i \in \mathcal{S}, j \in \mathcal{S}_i\}$ The conditional expectation of Y_{kl} given the random variables in the sample is:

$$E\left[Y_{kl}|\mathbf{Y}_{\mathcal{S}}\right] = \begin{cases} Y_{kl} & \text{if } i \in \mathcal{S} \text{ and } j \in \mathcal{S}_i \\ E\left[Y_{kl}|\mathbf{Y}_{\mathcal{S}_i}\right] & \text{if } i \in \mathcal{S} \text{ and } j \notin \mathcal{S}_i \\ \mu & \text{if } i \notin \mathcal{S} \end{cases}.$$

If psu k is not in the sample, then under model M1, its members are independent of all random variables in the sample and hence the expected value for all random variables in unsampled psus is μ. Thus, the only random variables where there is partial information from the sample are unsampled ssus within sampled psus. Under standard results for the multivariate normal distribution, when $j \notin \mathcal{S}_i$,

$$E\left[Y_{ij}|\mathbf{Y}_{\mathcal{S}_i}\right] = \mu + \sigma_A^2\mathbf{1}_{m_i}^T V_i^{-1}\left(\mathbf{Y}_{\mathcal{S}_i} - \mathbf{1}_{m_i}\mu\right),$$

where $V_i = \sigma_A^2 \mathbf{1}_{m_i}\mathbf{1}_{m_i}^T + \sigma^2 \mathbf{I}_{m_i}$. The inverse of V_i is

$$V_i^{-1} = \frac{1}{\sigma^2}\mathbf{I}_{m_i} - \frac{\sigma_A^2}{\sigma^4(1+m_i\sigma_A^2/\sigma^2)}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T = \frac{1}{\sigma^2}\mathbf{I}_{m_i} - \frac{\sigma_A^2}{\sigma^2(\sigma^2+m_i\sigma_A^2)}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T.$$

Consequently,

$$\begin{aligned}
E\left[Y_{ij}|\mathbf{Y}_{\mathcal{S}_i}\right] &= \mu + \sigma_A^2 \mathbf{1}_{m_i}^T V_i^{-1}\left(\mathbf{Y}_{\mathcal{S}_i} - \mathbf{1}_{m_i}\mu\right) \\
&= \mu + \sigma_A^2 \mathbf{1}_{m_i}^T \left(\frac{1}{\sigma^2}\mathbf{I}_{m_i} - \frac{\sigma_A^2}{\sigma^2(\sigma^2+m_i\sigma_A^2)}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\right)\left(\mathbf{Y}_{\mathcal{S}_i} - \mathbf{1}_{m_i}\mu\right) \\
&= \mu + \frac{m_i\sigma_A^2}{\sigma^2}(\bar{Y}_{\mathcal{S}_i} - \mu) - \frac{m_i^2\sigma_A^4}{\sigma^2(\sigma^2+m_i\sigma_A^2)}(\bar{Y}_{\mathcal{S}_i} - \mu) \\
&= \mu + \frac{m_i\sigma_A^2}{\sigma^2(\sigma^2+m_i\sigma_A^2)}\left[\sigma^2 + m_i\sigma_A^2 - m_i\sigma_A^2\right](\bar{Y}_{\mathcal{S}_i} - \mu) \\
&= \mu + \frac{m_i\sigma_A^2}{\sigma^2+m_i\sigma_A^2}(\bar{Y}_{\mathcal{S}_i} - \mu) \\
&= \mu + \alpha_i\rho(\bar{Y}_{\mathcal{S}_i} - \mu).
\end{aligned}$$

In model M1, $\rho = \sigma_A^2/(\sigma^2+\sigma_A^2)$, so we can write

$$\sigma_A^2 = \sigma^2\frac{\rho}{1-\rho}$$

and

$$\frac{m_i\sigma_A^2}{\sigma^2+m_i\sigma_A^2} = \frac{m_i\sigma^2\rho/(1-\rho)}{\sigma^2+m_i\sigma^2\rho/(1-\rho)} = \frac{m_i\rho}{1-\rho+m_i\rho} = \rho\alpha_i.$$

Also,

$$\begin{aligned}
\hat{\mu} &= \left(\sum_{i\in\mathcal{S}} \mathbf{1}_{m_i}^T V_i^{-1}\mathbf{1}_{m_i}\right)^{-1}\sum_{i\in\mathcal{S}} \mathbf{1}_{m_i}^T V_i^{-1}\mathbf{Y}_{\mathcal{S}_i} \\
&= \left(\sum_{i\in\mathcal{S}} \frac{m_i}{\sigma^2+m_i\sigma_A^2}\right)^{-1}\sum_{i\in\mathcal{S}} \frac{m_i}{\sigma^2+m_i\sigma_A^2}\bar{Y}_{\mathcal{S}_i} \\
&= \left(\sum_{i\in\mathcal{S}} \frac{m_i}{1+(m_i-1)\rho}\right)^{-1}\sum_{i\in\mathcal{S}} \frac{m_i}{1+(m_i-1)\rho}\bar{Y}_{\mathcal{S}_i} \\
&= \left(\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} \frac{1}{1+(m_i-1)\rho}\right)^{-1}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} \frac{m_i}{1+(m_i-1)\rho}Y_{ij}.
\end{aligned}$$

5.36 (a) Let's show that the predictor in (5.43) gives the constants c_{ij}.

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

$$
\begin{aligned}
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} Y_{ij} + \sum_{i\in\mathcal{S}}(M_i - m_i)\rho\frac{\alpha_i}{m_i}\sum_{l\in\mathcal{S}_i} Y_{il} + \left(\frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\frac{\alpha_i}{m_i}Y_{ij}}{\sum_{k\in\mathcal{S}}\alpha_k}\right)\left(\sum_{k\in\mathcal{S}}(M_k - m_k)(1-\rho\alpha_k) + \sum_{k\notin\mathcal{S}} M_k\right)\\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} Y_{ij}\left[1 + (M_i - m_i)\rho\frac{\alpha_i}{m_i} + \left(\frac{\alpha_i}{m_i}\frac{1}{\sum_{k\in\mathcal{S}}\alpha_k}\right)\left(\sum_{k\in\mathcal{S}}(M_k - m_k)(1-\rho\alpha_k) + \sum_{k\notin\mathcal{S}} M_k\right)\right]\\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} Y_{ij}\left[1 + (M_i - m_i)\rho\frac{\alpha_i}{m_i} + \left(\frac{\alpha_i}{m_i}\frac{1}{\sum_{k\in\mathcal{S}}\alpha_k}\right)\left(M_0 - \sum_{k\in\mathcal{S}} M_k\rho\alpha_k - \sum_{k\in\mathcal{S}} m_k(1-\rho\alpha_k)\right)\right]\\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} Y_{ij}\frac{\alpha_i}{m_i}\left[1 + \rho(m_i - 1) + (M_i - m_i)\rho + \left(\frac{1}{\sum_{k\in\mathcal{S}}\alpha_k}\right)\left(M_0 - \sum_{k\in\mathcal{S}} M_k\rho\alpha_k - \sum_{k\in\mathcal{S}} m_k(1-\rho\alpha_k)\right)\right]\\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} Y_{ij}\frac{\alpha_i}{m_i}\left[1 - \rho + \rho M_i + \left(\frac{1}{\sum_{k\in\mathcal{S}}\alpha_k}\right)\left(M_0 - \sum_{k\in\mathcal{S}} M_k\rho\alpha_k - \sum_{k\in\mathcal{S}}(1-\rho)\alpha_k\right)\right]\\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} Y_{ij}\frac{\alpha_i}{m_i}\left[\rho M_i + \left(\frac{1}{\sum_{k\in\mathcal{S}}\alpha_k}\right)\left(M_0 - \sum_{k\in\mathcal{S}} M_k\rho\alpha_k\right)\right]
\end{aligned}
$$

(b) The estimator will be unbiased if $\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} c_{ij} = M_0$, so let's verify that.

$$
\begin{aligned}
\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\frac{\alpha_i}{m_i}\left[\rho M_i + \frac{M_0 - \rho\sum_{k\in\mathcal{S}}\alpha_k M_k}{\sum_{k\in\mathcal{S}}\alpha_k}\right] &= \sum_{i\in\mathcal{S}}\alpha_i\left[\rho M_i + \frac{M_0 - \rho\sum_{k\in\mathcal{S}}\alpha_k M_k}{\sum_{k\in\mathcal{S}}\alpha_k}\right]\\
&= M_0 + \frac{\sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S}}\alpha_i\left[\rho M_i\alpha_k - \rho\alpha_k M_k\right]}{\sum_{k\in\mathcal{S}}\alpha_k}\\
&= M_0.
\end{aligned}
$$

(c) Let $\hat{T}_2$ be another linear unbiased predictor of T. Because it is unbiased, it must be of the form

$$\hat{T}_2 = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(c_{ij} + a_{ij})Y_{ij},$$

where

$$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} a_{ij} = 0.$$

From (5.42),

$$V_{\mathrm{M1}}[\hat{T}_2 - T] = \sigma_A^2\left[\sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}_i}(c_{ij} + a_{ij}) - M_i\right)^2 + \sum_{i\notin\mathcal{S}} M_i^2\right]$$

$$
\begin{aligned}
&+ \sigma^2 \left[\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}_i} \{(c_{ij}+a_{ij})^2 - 2c_{ij} - 2a_{ij}\} + M_0 \right] \\
&= \sigma_A^2 \sum_{i\in\mathcal{S}} \left[\left(\sum_{j\in\mathcal{S}_i} c_{ij} - M_i \right)^2 + \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right)^2 + 2 \left(\sum_{j\in\mathcal{S}_i} c_{ij} - M_i \right) \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) + \sum_{i\notin\mathcal{S}} M_i^2 \right] \\
&\quad + \sigma^2 \left[\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}_i} \{c_{ij}^2 + 2c_{ij}a_{ij} + a_{ij}^2 - 2c_{ij} - 2a_{ij}\} + M_0 \right] \\
&= V_{\mathrm{M1}}(\hat{T}_{\mathrm{BLUP}}) + \sigma_A^2 \left[\sum_{i\in\mathcal{S}} \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right)^2 + 2 \left(\sum_{j\in\mathcal{S}_i} c_{ij} - M_i \right) \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) \right] \\
&\quad + \sigma^2 \left[\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}_i} \{a_{ij}^2 + 2c_{ij}a_{ij} - 2a_{ij}\} \right] \qquad (5.1)
\end{aligned}
$$

Now let's look at the terms involving c_{ij} in (SM5.1). We know from part (a) that

$$
c_{ij} = \frac{\alpha_i}{m_i} \left[\rho M_i + \frac{M_0 - \rho \sum_{k\in\mathcal{S}} \alpha_k M_k}{\sum_{k\in\mathcal{S}} \alpha_k} \right] = \frac{1}{1+\rho(m_i-1)} \left[\rho M_i + K \right]
$$

where

$$
K = \frac{M_0 - \rho \sum_{k\in\mathcal{S}} \alpha_k M_k}{\sum_{k\in\mathcal{S}} \alpha_k}
$$

is a constant that does not depend on i. Consequently, noting that $\rho = \sigma_A^2/(\sigma_A^2+\sigma^2)$,

$$
\begin{aligned}
&\sigma_A^2 \sum_{i\in\mathcal{S}} \left(\sum_{j\in\mathcal{S}_i} c_{ij} - M_i \right) \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) + \sigma^2 \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}_i} c_{ij}a_{ij} \\
&= \sigma_A^2 \sum_{i\in\mathcal{S}} \left(\frac{m_i}{1+\rho(m_i-1)} \left[\rho M_i + K \right] - M_i \right) \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) + \sigma^2 \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}_i} \frac{\rho M_i + K}{1+\rho(m_i-1)} a_{ij} \\
&= (\sigma_A^2+\sigma^2) \sum_{i\in\mathcal{S}} \left(\frac{\rho m_i}{1+\rho(m_i-1)} \left[\rho M_i + K \right] - \rho M_i + (1-\rho) \frac{\rho M_i + K}{1+\rho(m_i-1)} \right) \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) \\
&= (\sigma_A^2+\sigma^2) \sum_{i\in\mathcal{S}} \left(\frac{\rho m_i + 1 - \rho}{1+\rho(m_i-1)} \left[\rho M_i + K \right] - \rho M_i \right) \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) \\
&= (\sigma_A^2+\sigma^2) \sum_{i\in\mathcal{S}} K \left(\sum_{j\in\mathcal{S}_i} a_{ij} \right) \\
&= 0
\end{aligned}
$$

because $\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} a_{ij} = 0$. Thus, returning to (SM5.1),

$$\begin{aligned}
V_{\mathrm{M1}}[\hat{T}_2 - T] &= V_{\mathrm{M1}}(\hat{T}_{\mathrm{BLUP}}) + \sigma_A^2 \left[\sum_{i\in\mathcal{S}} \left(\sum_{j\in\mathcal{S}_i} a_{ij}\right)^2 + 2\left(\sum_{j\in\mathcal{S}_i} c_{ij} - M_i\right)\left(\sum_{j\in\mathcal{S}_i} a_{ij}\right)\right] \\
&\quad + \sigma^2 \left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} \{a_{ij}^2 + 2c_{ij}a_{ij} - 2a_{ij}\}\right] \\
&= V_{\mathrm{M1}}(\hat{T}_{\mathrm{BLUP}}) + \sigma_A^2 \left[\sum_{i\in\mathcal{S}} \left(\sum_{j\in\mathcal{S}_i} a_{ij}\right)^2\right] + \sigma^2 \left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} a_{ij}^2\right] \\
&\geq V_{\mathrm{M1}}(\hat{T}_{\mathrm{BLUP}}).
\end{aligned}$$

5.37 (a) Examples that are described by this model are harder to come by in practice, but let's construct one based on the principle that tasks expand to fill up the allotted time. All students at Idyllic College have 100 hours available for writing term papers, but an individual student may have from one to five papers assigned. It would never occur to an Idyllic student to finish a paper quickly and relax in the extra time, so a student with one paper spends all 100 hours on the paper, a student with two papers spends 50 on each, and so on. Thus, the expected total amount of time spent writing term papers, $E[T_i]$, is 100 for each student, although the numbers of papers assigned (M_i) vary.

(b)

$$\begin{aligned}
E_{\mathrm{M2}}[\hat{T} - T] &= E_{\mathrm{M2}}\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}Y_{ij} - \sum_{i=1}^{N}\sum_{j=1}^{M_i} Y_{ij}\right] \\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij} E_{\mathrm{M2}}\left[B_i + \varepsilon_{ij}\right] - \sum_{i=1}^{N}\sum_{j=1}^{M_i} E_{\mathrm{M2}}\left[B_i + \varepsilon_{ij}\right] \\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}\frac{\mu}{M_i} - \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{\mu}{M_i} \\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} b_{ij}\frac{\mu}{M_i} - N\mu.
\end{aligned}$$

For estimator $\hat{T}_{\mathrm{unb}}$, $b_{ij} = (NM_i)/(nm_i)$ and

$$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} \frac{NM_i}{nm_i}\frac{\mu}{M_i} - N\mu = N\mu - N\mu = 0.$$

(c) This is done similarly to Exercise 5.34.

$$
\begin{aligned}
V_{\mathrm{M2}}\left[\hat{T}-T\right] &= V_{\mathrm{M2}}\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(b_{ij}-1)Y_{ij}-\sum_{i\in\mathcal{S}}\sum_{j\notin\mathcal{S}_i}Y_{ij}-\sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i}Y_{ij}\right] \\
&= V_{\mathrm{M2}}\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(b_{ij}-1)(B_i+\varepsilon_{ij})-\sum_{i\in\mathcal{S}}\sum_{j\notin\mathcal{S}_i}(B_i+\varepsilon_{ij})-\sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i}(B_i+\varepsilon_{ij})\right] \\
&= V_{\mathrm{M2}}\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(b_{ij}-1)B_i-\sum_{i\in\mathcal{S}}\sum_{j\notin\mathcal{S}_i}B_i-\sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i}B_i\right] \\
&\quad + V_{\mathrm{M2}}\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(b_{ij}-1)\varepsilon_{ij}-\sum_{i\in\mathcal{S}}\sum_{j\notin\mathcal{S}_i}\varepsilon_{ij}-\sum_{i\notin\mathcal{S}}\sum_{j=1}^{M_i}\varepsilon_{ij}\right] \\
&= V_{\mathrm{M2}}\left[\sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}_i}b_{ij}-m_i\right)B_i-\sum_{i\in\mathcal{S}}(M_i-m_i)B_i-\sum_{i\notin\mathcal{S}}M_iB_i\right] \\
&\quad + \sigma^2\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(b_{ij}-1)^2+\sum_{i\in\mathcal{S}}(M_i-m_i)+\sum_{i\notin\mathcal{S}}M_i\right] \\
&= \sigma_B^2\left[\sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}_i}b_{ij}-M_i\right)^2\frac{1}{M_i^2}+\sum_{i\notin\mathcal{S}}\frac{1}{M_i}\right] \\
&\quad + \sigma^2\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}(b_{ij}-1)^2+\sum_{i\in\mathcal{S}}(M_i-m_i)+\sum_{i\notin\mathcal{S}}M_i\right].
\end{aligned}
$$

5.38 The puppy homes certainly follow Model M1: All puppies have four legs, so $Y_{ij}=\mu=4$ for all i and j. Consequently, $\sigma_A^2=\sigma^2=0$. The model-based variance of the estimate $\hat{T}_{\mathrm{unb}}$ is therefore 0, no matter which puppy home and puppies are chosen. If Puppy Palace is selected for the sample, the bias under the model in (5.38) is $4(2\times 30-40)=80$; if Dog's Life is selected, the bias is $4(2\times 10-40)=-80$.

In the design-based approach in Example 5.9, the unbiased estimator $\hat{t}_{\mathrm{unb}}$ had a large design-based variance. The large variance in the design-based approach thus becomes a bias when a model-based approach is adopted. It is not surprising that $\hat{T}_{\mathrm{unb}}$ performs poorly for the puppy homes; it is a poor estimator for a model that describes the situation well.

The model-based bias and variance for $\hat{T}_r$, though, are both zero.

5.39 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 (5.42),

$$\begin{aligned} V_{M1}[\hat{T}_r - T] &= \sigma_A^2 \left[\sum_{i \in \mathcal{S}} \left(\sum_{j \in \mathcal{S}_i} \frac{M_0 M_i}{m_i \sum_{k \in \mathcal{S}} M_k} - M_i \right)^2 + \sum_{i \notin \mathcal{S}} M_i^2 \right] \\ &\quad + \sigma^2 \left[\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \left(\left(\frac{M_0 M_i}{m_i \sum_{k \in \mathcal{S}} M_k} \right)^2 - 2 \frac{M_0 M_i}{m_i \sum_{k \in \mathcal{S}} M_k} \right) + M_0 \right] \\ &= \sigma_A^2 \left[\sum_{i \in \mathcal{S}} M_i^2 \left(\frac{M_0}{\sum_{k \in \mathcal{S}} M_k} - 1 \right)^2 + \sum_{i \notin \mathcal{S}} M_i^2 \right] \\ &\quad + \sigma^2 \sum_{i \in \mathcal{S}} \left[\frac{M_0^2 M_i^2}{m_i (\sum_{k \in \mathcal{S}} M_k)^2} - 2 \frac{M_0 M_i}{\sum_{k \in \mathcal{S}} M_k} \right] + \sigma^2 M_0, \end{aligned}$$

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

6.1 Answers will vary.

6.2 (a) Instead of creating columns of cumulative M_i range as in Example 6.2, we create columns for the cumulative ψ_i range. Then draw 10 random numbers between 0 and 1 to select the psu's with replacement. Table SM6.1 gives the cumulative ψ_i range for each psu.

(Note: the numbers in the "Cumulative ψ_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 psus 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	ψ_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

Table 6.1: Cumulative ψ_i range for each psu.

psu	ψ_i	Cumulative ψ_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

$$E[\hat{t}_\psi] = \frac{1}{16}(1200) + \frac{2}{16}(600) + \frac{3}{16}(400) + \frac{10}{16}(120) = 300.$$
$$V[\hat{t}_\psi] = \frac{1}{16}(810{,}000) + \cdots + \frac{10}{16}(32{,}400) = 84{,}000.$$

6.4

Store	ψ_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,

$$E[\hat{t}_\psi] = \frac{7}{16}(25.14) + \frac{3}{16}(106.67) + \frac{3}{16}(128) + \frac{3}{16}(1306.67) = 300.$$

Using (6.4),

$$V[\hat{t}_\psi] = \frac{7}{16}(75546.4) + \frac{3}{16}(37377.8) + \frac{3}{16}(29584) + \frac{3}{16}(1013377.8) = 235615.2.$$

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 ψ_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/ψ_1	t_2/ψ_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) Let's write an R function, using (6.4) to find the variance.

```
psivar <- function(psii,ti) {
   if (any(psii<= 0)) return ("psii must be > 0")
   # normalize so sums to one
   psii <- psii/sum(psii)
   tpop <- sum(ti)
   varcontr <- psii*(ti/psii - tpop)^2
   sum(varcontr)
}

data(azcounties)
# calculate variance for unequal probabilities
psivar(azcounties$population,azcounties$housing)
## [1] 13620222734
# calculate variance for SRS
psivar(rep(1,nrow(azcounties)),azcounties$housing)
## [1] 386553128504
cor(azcounties$housing,azcounties$population)
## [1] 0.9829381
# look at ratio
psivar(rep(1,nrow(azcounties)),azcounties$housing)/psivar(azcounties$population,
    azcounties$housing)
## [1] 28.38082
```

We have $V(\hat{t}_\psi)$ =13620222734.

(b) Using $\psi_i = 1/13$ for each i, we get the SRS estimators with $V(\hat{t}_{SRS})$ =386,553,128,504. This is more than 28 times as high as the variance from the unequal-probability sample!

The unequal-probability sample is more efficient because t_i and M_i are highly correlated: the correlation is 0.983. This means that the quantity t_i/ψ_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

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(Nt_i - N\bar{t}\right)^2 = \frac{N^2}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(t_i - \bar{t}\right)^2.$$

6.8 We use (6.13) and (6.14), along with the following calculations from a spreadsheet.

Academic Unit	M_i	ψ_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       .   .
```

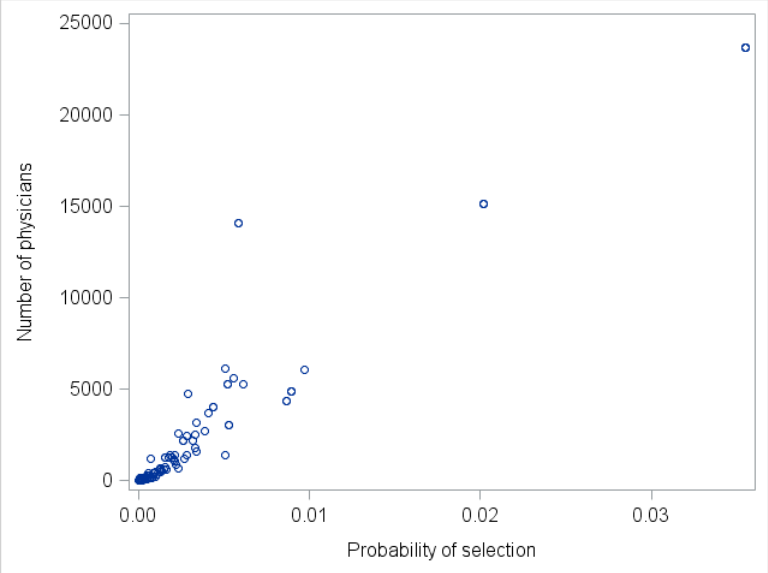
```
6       62      0.0768278       0       2       2       5
14c     65      0.0805452       1       0       0       3
19      62      0.0768278       4       1       0       0
21      61      0.0755886       2       2       3       1
11      41      0.0508055       2       5       12      3
;

proc surveymeans data=faculty mean clm sum clsum;
   cluster unit;
   weight sampwt;
   var y;
run;
```

6.9–6.11 Answers will vary.

6.12 (a) Figure SM6.1 shows the scatterplot. You would expect larger counties to have more physicians, so pps sampling should work well for estimating the total number of physicians in the United States.

Figure 6.1: Plot of t_i (number of physicians) versus ψ_i; there is a strong linear relationship between the variables, which indicates that pps sampling increases efficiency.



The sample was chosen using the cumulative size method; Table SM6.2 shows the sampled counties arranged alphabetically by state. The ψ_i's were calculated using $\psi_i = M_i/M_0$. The average of the t_i/ψ_i column is 570,304.3, the estimated total number of physicians in the United States. The standard error of the estimate is $414{,}012.3/\sqrt{100} = 41{,}401$. For comparison, the *County and City Data Book* lists a total of 532,638 physicians in the United States; a 95% CI using our estimate includes the true value.

Table 6.2: Sampled counties for Exercise 12.

State	County	Population Size, M_i	ψ_i	Number of Physicians, t_i	t_i/ψ_i
AL	Wilcox	13,672	0.00005360	4	74,627.72
AZ	Maricopa	2,209,567	0.00866233	4320	498,710.81
AZ	Maricopa	2,209,567	0.00866233	4320	498,710.81
AZ	Pinal	120,786	0.00047353	61	128,820.64
AR	Garland	76,100	0.00029834	131	439,095.36
⋮	⋮	⋮	⋮	⋮	⋮
VA	Chesterfield	225,225	0.00088297	181	204,990.72
WA	King	1,557,537	0.00610613	5280	864,704.59
WI	Lincoln	27,822	0.00010907	28	256,709.47
WI	Waukesha	320,306	0.00125572	687	547,096.42
		average			570,304.30
		std. dev.			414,012.30

(d) Here is code from SAS software

```
data statepop;
   set datalib.statepop;
   u = phys/psii;

proc means data=statepop mean std;
   var u;

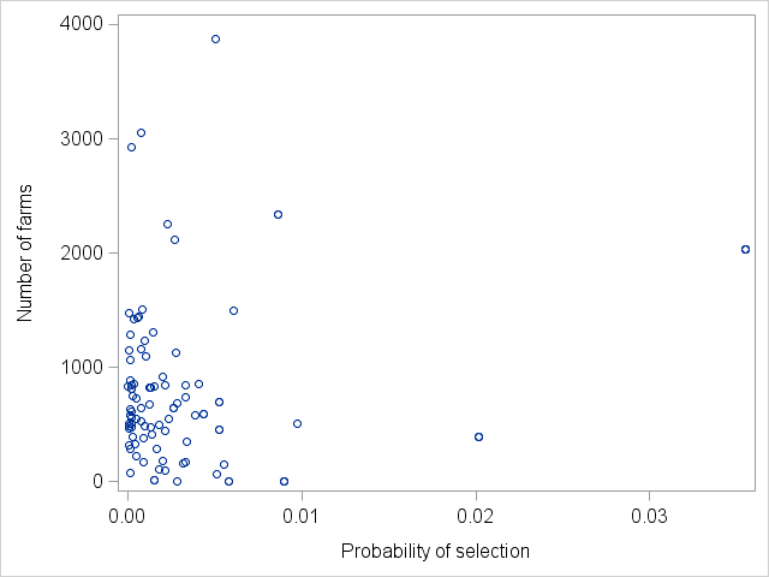
proc corr data=statepop;
   var psii phys numfarm veterans;

proc sgplot data=statepop;
   scatter x = psii y = phys;
   xaxis label = "Probability of selection";
   yaxis label = "Number of physicians";

/* Omit the total option because sampling is with replacement */
proc surveymeans data=statepop mean clm sum clsum;
   weight wt;
   var phys numfarm veterans;
run;
```

Here is the output. The sum of the weights is 2450.72.

Figure 6.2: Scatterplot for Exercise 6.13



Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum	95% CL for Sum	
phys	232.7089	48.8593	135.7615	329.6564	570304	41401	488155.3	652453.3
numfarm	773.7728	81.5264	612.0068	935.5389	1896300	367423	1167254.2	2625346.2
veterans	11390	1961.3672	7498.4196	15281.9758	27914180	1087453	25756437	30071923

6.13 (a) The correlation between ψ_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 Exercise 12. Figure SM6.2 shows the plot.

(b) As in Exercise 6.12, we form a new column t_i/ψ_i. The mean of this column is 1,896,300, and the standard deviation of the column is 3,674,225. Thus

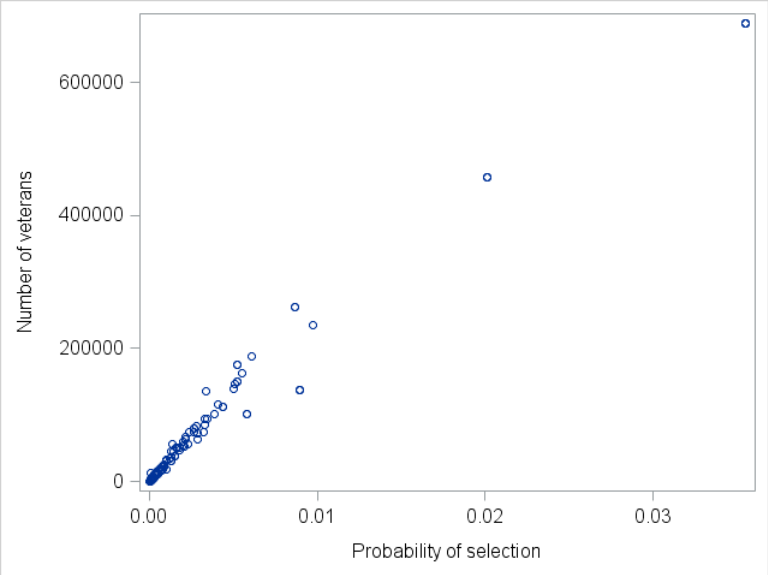
$$\hat{t}_\psi = 1{,}896{,}300 \text{ and } \mathrm{SE}\,[\hat{t}_\psi] = \frac{3{,}674{,}225}{\sqrt{100}} = 367{,}423.$$

A histogram of the t_i/ψ_i exhibits strong skewness, however, so a confidence interval using the normal approximation may not be appropriate.

See Exercise 6.12 for code and estimates.

6.14 (a) Corr (probability of selection, number of veterans) = 0.99, as can be seen in Figure SM6.3. We expect the unequal probability sampling to be very efficient here.

Figure 6.3: Plot of veterans versus probability of selection



(b)

$$\hat{t}_\psi = \frac{1}{100}\sum \frac{t_i}{\psi_i} = 27{,}914{,}180$$
$$\text{SE}\left[\hat{t}_\psi\right] = \frac{10874532}{\sqrt{100}} = 1{,}087{,}453$$

Note that Allen County, Ohio appears to be an outlier among the t_i/ψ_i: it has population 26,405 and 12,642 veterans.

6.15 Table SM6.3 gives the selection probability ψ_i and weight $w_i = 1/(10\psi_i)$ for each college in the sample, as well as the response t_i = number of biology faculty members. One would expect that the number of faculty members would be roughly proportional to the number of undergraduates, so that pps sampling would be a good choice for this response variable.

The estimated total number of biology faculty members for the stratum is $\hat{t}_\psi = \sum_{i\in\mathcal{W}} w_i t_i = 3237.18$ with

$$\hat{V}\left(\hat{t}_\psi\right) = \frac{1}{10}\frac{1}{9}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2 = 86434.9$$

and $\text{SE}(\hat{t}_\psi) = \sqrt{86434.9} = 294$. The 95% CI for the total, using a t distribution with 9 degrees of freedom, is [2572, 3902].

The ratio estimator for the mean number of biology faculty members per college is

$$\hat{\bar{t}} = \frac{\sum_{i\in\mathcal{W}} w_i t_i}{\sum_{i\in\mathcal{W}} w_i} = \frac{3237.2}{325.37} = 9.95.$$

Table 6.3: Sampled colleges for Exercise 6.15.

College	ψ_i	w_i	t_i	$w_i t_i$
University of Arkansas at Pine Bluff	0.00434	23.06	11	253.67
Augustana College	0.00445	22.50	15	337.43
University of Northwestern-St Paul	0.00320	31.29	14	438.10
Mount Vernon Nazarene University	0.00294	34.05	5	170.23
The College of Wooster	0.00352	28.43	15	426.47
Bryn Mawr College	0.00238	42.05	10	420.48
Moravian College	0.00362	27.60	10	276.03
Waynesburg University	0.00232	43.03	5	215.17
McMurry University	0.00190	52.68	7	368.73
Carthage College	0.00484	20.68	16	330.88
Sum		325.37		3237.18

Letting s_e^2 be the sample variance of the residuals $e_i = w_i t_i - w_i \hat{\bar{t}}$, by (4.12), without the fpc,

$$\text{SE}\left(\hat{\bar{t}}\right) = \sqrt{\frac{s_e^2}{n\bar{w}^2}} = \sqrt{\frac{18193.38}{(10)(32.537)^2}} = 1.311.$$

6.16 The estimated total number of math faculty is 2916.9, with 95% CI [2034.4, 3799.3]. The estimated total number of psychology faculty is 2343.6, with 95% CI [1468.7, 3218.6].

6.17 The estimated total number of female undergraduates is 350673 with 95% CI [294006.9, 407339.2]. The correlation between number of female undergraduates and ψ_i is 0.82. Pps sampling is a good idea for this sample. The scatterplot is shown in Figure SM6.4.

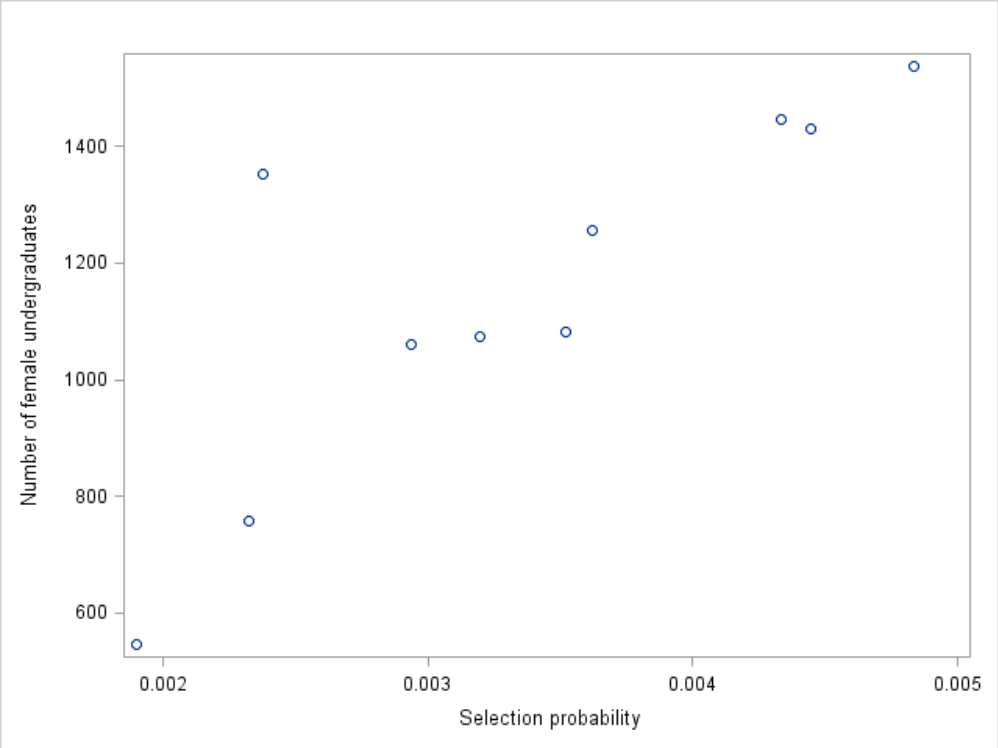
6.18 (a) We use (6.28) and (6.29) to calculate the variances. We have

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 6.8, and calculating

$$\hat{V}(\hat{t}_i) = M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i},$$

Figure 6.4: Scatterplot for Exercise 6.17



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.

```
                              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.

6.19 (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,

$$
\begin{aligned}
P(\text{select psu } i) &= P(\text{select psu } i \text{ with first pair of random numbers}) \\
&\quad + P(\text{select psu } i \text{ with second pair}) \\
&\quad + \cdots \\
&= P(U_1 = i \quad \text{and} \quad V_1 \le M_i) \\
&\quad + P(U_2 = i, V_2 \le M_i) P\Big(\bigcup_{j=1}^{N} \{U_1 = j, V_2 > M_j\} \Big) \\
&\quad + \cdots + P(U_k = i, V_k \le M_i) \prod_{l=1}^{k-1} P\Big(\bigcup_{j=1}^{N} \{U_l = j, V_l > M_j\} \Big) \\
&= \frac{1}{N}\frac{M_i}{J} + \frac{1}{N}\frac{M_i}{J}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{J - M_j}{J}\right] + \cdots + \frac{1}{N}\frac{M_i}{J}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{J - M_j}{J}\right]^{k-1} + \cdots \\
&= \frac{1}{N}\frac{M_i}{J}\sum_{k=0}^{\infty}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{J - M_j}{J}\right]^{k} \\
&= \frac{1}{N}\frac{M_i}{J}\frac{1}{1 - \sum_{j=1}^{N}(J - M_j)/(JN)} \\
&= M_i \Big/ \sum_{j=1}^{N} M_j.
\end{aligned}
$$

(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,

$$E[X] = nNJ \Big/ \sum_{i=1}^{N} M_i.$$

6.20 The random variables $Q_1, \ldots, Q_N$ have a joint multinomial distribution with probabilities $\psi_1, \ldots, \psi_N$. Consequently,

$$\begin{aligned} &\sum_{i=1}^{n} Q_i = n, \\ &E[Q_i] = n\psi_i, \\ &V[Q_i] = n\psi_i(1-\psi_i), \end{aligned}$$

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

$$\begin{aligned} V(\hat{t}_\psi) &= V\left[E(\hat{t}_\psi \mid Q_1, \ldots, Q_N)\right] + E[V(\hat{t}_\psi \mid Q_1, \ldots, Q_N)] \\ &= V\Big[\frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}\Big] + E\Big[V\Big(\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i}\,\Big|\, Q_1, \ldots, Q_N\Big)\Big]. \end{aligned}$$

Looking at the two terms separately,

$$\begin{aligned} V\Big[\frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}\Big] &= \mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}, \frac{1}{n}\sum_{k=1}^{N} Q_k \frac{t_k}{\psi_k}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{t_i}{\psi_i}\frac{t_k}{\psi_k}\,\mathrm{Cov}\,(Q_i, Q_k) \\ &= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{t_i}{\psi_i}\right)^2 n\psi_i(1-\psi_i) + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\frac{t_i}{\psi_i}\frac{t_k}{\psi_k}(-n\psi_i\psi_k) \\ &= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{t_i}{\psi_i}\right)^2[\psi_i(1-\psi_i)+\psi_i^2] - \frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{N} t_i t_k \\ &= \frac{1}{n}\sum_{i=1}^{N}\psi_i\Big(\frac{t_i}{\psi_i} - t\Big)^2. \end{aligned}$$

$$\begin{aligned} E\Big[V\Big(\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i}\mid Q_1, \ldots, Q_N\Big)\Big] &= E\Big[\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{V(\hat{t}_{ij})}{\psi_i^2}\Big] \\ &= E\Big[\frac{1}{n^2}\sum_{i=1}^{N} Q_i\frac{V_i}{\psi_i^2}\Big] \\ &= \frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{\psi_i}. \end{aligned}$$

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,

$$V[\hat{t}_\psi] = \frac{1}{n}\sum_{i=1}^{N}\psi_i\Big(\frac{t_i}{\psi_i} - t\Big)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{\psi_i}.$$

(c) To show that (6.14) is an unbiased estimator of the variance, note that

$$E[\hat{V}(\hat{t}_\psi)] \;=\; E\Big[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{(\hat{t}_{ij}/\psi_i - \hat{t}_\psi)^2}{n-1}\Big]$$

$$
\begin{aligned}
&= E\Big\{\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{1}{n-1}\Big[\Big(\frac{\hat{t}_{ij}}{\psi_i}\Big)^2 - 2\frac{\hat{t}_{ij}}{\psi_i}\hat{t}_\psi + \hat{t}_\psi^2\Big]\Big\}\\
&= E\Big[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{1}{n-1}\Big(\frac{\hat{t}_{ij}}{\psi_i}\Big)^2 - \frac{\hat{t}_\psi^2}{n-1}\Big]\\
&= E\Big\{\frac{1}{n}E\Big[\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{1}{n-1}\Big(\frac{\hat{t}_{ij}}{\psi_i}\Big)^2\Big|Q_1,\ldots,Q_N\Big]\Big\} - \frac{1}{n-1}[t^2 + V(\hat{t}_\psi)]\\
&= E\Big[\frac{1}{n}\sum_{i=1}^{N}\frac{Q_i}{n-1}\frac{t_i^2+V_i}{\psi_i^2}\Big] - \frac{1}{n-1}[t^2 + V(\hat{t}_\psi)]\\
&= \frac{1}{n-1}\sum_{i=1}^{N}\frac{t_i^2+V_i}{\psi_i} - \frac{1}{n-1}[t^2 + V(\hat{t}_\psi)]\\
&= \frac{1}{n-1}\Big(\sum_{i=1}^{N}\psi_i\frac{t_i^2}{\psi_i^2} - t^2\Big) + \frac{1}{n-1}\Big(\sum_{i=1}^{N}\frac{V_i}{\psi_i} - V(\hat{t}_\psi)\Big)\\
&= \frac{1}{n-1}\Big[\sum_{i=1}^{N}\psi_i\Big(\frac{t_i}{\psi_i} - t\Big)^2 + \sum_{i=1}^{N}\frac{V_i}{\psi_i}\Big] - \frac{1}{n-1}V(\hat{t}_\psi)\\
&= \frac{n}{n-1}V(\hat{t}_\psi) - \frac{1}{n-1}V(\hat{t}_\psi) = V(\hat{t}_\psi).
\end{aligned}
$$

6.21 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,

$$
\begin{aligned}
&\frac{1}{2}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2\\
&\quad= \frac{1}{2}\frac{1}{\pi_1^2}\frac{\pi_1^2-\pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}(t_i - t_k)^2\\
&\quad= \frac{1}{2}\frac{1}{\pi_1^2}\frac{\pi_1^2-\pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}(t_i^2 + t_k^2 - 2t_it_k)\\
&\quad= \frac{1}{\pi_1^2}\frac{\pi_1^2-\pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}(n-1)t_i^2 - \frac{1}{\pi_1^2}\frac{\pi_1^2-\pi_{12}}{\pi_{12}}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}t_it_k.
\end{aligned}
$$

But in an SRS,

$$
\frac{1}{\pi_1^2}\frac{\pi_1^2-\pi_{12}}{\pi_{12}}(n-1) = \left(\frac{N}{n}\right)^2\frac{\dfrac{n}{N} - \dfrac{n-1}{N-1}}{\dfrac{n-1}{N-1}}(n-1) = \left(\frac{N}{n}\right)^2\left(1 - \frac{n}{N}\right),
$$

which proves the result.

6.22 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

$$\hat{t}_{\text{HT}} = \sum_{i=1}^{H} \hat{t}_i = \sum_{i=1}^{H} N_i \bar{y}_i.$$

Since all strata are sampled, either (6.26) or (6.27) gives $V(\hat{t}_{\text{HT}}) = \sum_{i=1}^{H} V(\hat{t}_i)$.

Result (3.6) follows from (6.29) similarly.

6.23 (a) We use the method in Example 6.8 to calculate the probabilities.

					π_{jk}					
Unit	1	2	3	4	5	6	7	8	9	π_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
π_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), $V(\hat{t}_{\text{HT}}) = 1643$. Using (6.46) (we only need the first term), $V_{WR}(\hat{t}_\psi) = 1867$.

6.24 (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,

$$\begin{aligned}
\mathrm{Cov}\,(\hat{t}_x, \hat{t}_y) &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_{ix} t_{iy} + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} (\pi_{ik} - \pi_i \pi_k) \frac{t_{ix}}{\pi_i} \frac{t_{ky}}{\pi_k} \\
&= \sum_{i=1}^{N} \frac{N}{n}\left(1-\frac{n}{N}\right) t_{ix} t_{iy} + \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} \left[\frac{n}{N}\frac{(n-1)}{(N-1)} - \left(\frac{n}{N}\right)^2\right] t_{ix} t_{iy} \\
&= \sum_{i=1}^{N} \frac{N}{n}\left(1-\frac{n}{N}\right) t_{ix} t_{iy} + \frac{N}{n} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} \left[\frac{(n-1)}{(N-1)} - \frac{n}{N}\right] t_{ix} t_{ky} \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right) \sum_{i=1}^{N} t_{ix} t_{iy} - \frac{N}{n} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} \frac{N-n}{N(N-1)} t_{ix} t_{ky} \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right) \left[\sum_{i=1}^{N} t_{ix} t_{iy} - \frac{1}{N-1} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} t_{ix} t_{ky}\right] \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right) \left[\sum_{i=1}^{N} t_{ix} t_{iy} - \frac{1}{N-1} \sum_{i=1}^{N} \sum_{k=1}^{N} t_{ix} t_{ky} + \frac{1}{N-1} \sum_{i=1}^{N} t_{ix} t_{iy}\right] \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right) \left[\frac{N}{N-1} \sum_{i=1}^{N} t_{ix} t_{iy} - \frac{1}{N-1} \sum_{i=1}^{N} \sum_{k=1}^{N} t_{ix} t_{ky}\right] \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right) \frac{1}{N-1} \left[\sum_{i=1}^{N} t_{ix} t_{iy} - \frac{t_x t_y}{N}\right]
\end{aligned}$$

6.25 *Comparing domains.* 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,

$$\begin{aligned}
&(NM)^2 \mathrm{Cov}\left[\left(\hat{\bar{u}} - \frac{t_u}{t_x}\hat{\bar{x}}\right),\ \left\{\hat{\bar{y}} - \hat{\bar{u}} - \frac{t_y - t_u}{N - t_x}(1 - \hat{\bar{x}})\right\}\right] \\
&= \mathrm{Cov}\left[\left(\hat{t}_u - \frac{t_u}{t_x}\hat{t}_x\right),\ \left\{\hat{t}_y - \hat{t}_u - \frac{t_y - t_u}{N - t_x}(NM - \hat{t}_x)\right\}\right] \\
&= \mathrm{Cov}\left[\hat{t}_u, \hat{t}_y - \hat{t}_u + \frac{t_y - t_u}{N - t_x}\hat{t}_x\right] - \frac{t_u}{t_x}\mathrm{Cov}\left[\hat{t}_x, \hat{t}_y - \hat{t}_u + \frac{t_y - t_u}{N - t_x}\hat{t}_x\right] \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N}\left\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x} t_{iu} t_{ix}\right.\right.
\end{aligned}$$

$$
\begin{aligned}
& \qquad\qquad -\frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\Bigg\}\Bigg] \\
& -\frac{N}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\Bigg[t_u t_y - t_u^2 + \frac{t_y - t_u}{N - t_x}t_u t_x \\
& \qquad\qquad -\frac{t_u}{t_x}\left(t_x t_y - t_x t_u + \frac{t_y - t_u}{N - t_x}t_x^2\right)\Bigg] \\
= \quad & \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\Bigg[\sum_{i=1}^{N}\Bigg\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x}t_{iu}t_{ix} \\
& \qquad\qquad -\frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\Bigg\}\Bigg].
\end{aligned}
$$

(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

$$
\begin{aligned}
& \sum_{i=1}^{N}\left\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x}t_{iu}t_{ix} - \frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\right\} \\
= \quad & \sum_{i\in\text{domain 1}}\left\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x}t_{iu}t_{ix} - \frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\right\} \\
& + \sum_{i\in\text{domain 2}}\left\{t_{iu}t_{iy} - t_{iu}^2 + \frac{t_y - t_u}{N - t_x}t_{iu}t_{ix} - \frac{t_u}{t_x}\left(t_{ix}t_{iy} - t_{ix}t_{iu} + \frac{t_y - t_u}{N - t_x}t_{ix}^2\right)\right\} \\
= \quad & \sum_{i\in\text{domain 1}}\left\{t_{iy}t_{iy} - t_{iy}^2 + \frac{t_y - t_u}{N - t_x}t_{iy}M - \frac{t_u}{t_x}\left(Mt_{iy} - Mt_{iy} + \frac{t_y - t_u}{N - t_x}M^2\right)\right\} \\
= \quad & \sum_{i\in\text{domain 1}}\left\{\frac{t_y - t_u}{N - t_x}t_{iy}M - \frac{t_u}{t_x}\frac{t_y - t_u}{N - t_x}M^2\right\} \\
= \quad & \frac{t_y - t_u}{N - t_x}t_u M - N\frac{t_x}{NM}\frac{t_u}{t_x}\frac{t_y - t_u}{N - t_x}M^2 \\
= \quad & 0.
\end{aligned}
$$

(c) Almost any example will work, as long as some psus have units from both domains.

6.26 *Indirect sampling.* (a) Since $E(Z_i) = \pi_i$,

$$E(\hat{t}_y) = \sum_{i=1}^{N} \frac{1}{\pi_i} u_i E(Z_i) = \sum_{i=1}^{N} u_i = \sum_{i=1}^{N} \sum_{k=1}^{M} \frac{\ell_{ik} y_k}{L_k} = \sum_{k=1}^{M} y_k.$$

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 $k \notin \mathcal{S}^B$.

$$w_k^* = \frac{\sum_{i=1}^{N} \frac{Z_i}{\pi_i} \ell_{ik}}{\sum_{i=1}^{N} \ell_{ik}}.$$

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 $\mathcal{U}^B$		
ℓ_{ik}		1	2	u_i
Unit i	1	1	0	4/2 = 2
from $\mathcal{U}^A$	2	1	1	4/2 + 6/2 = 5
	3	0	1	6/2 = 3

Here are the three SRSs from $\mathcal{U}^A$:

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;
   set datalib.wtshare;
   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;
```

6.27 *Poisson sampling*

(a) Because the random variables Z_i are independent,

$$V\left(\sum_{i=1}^{N} Z_i\right) = \sum_{i=1}^{N} V(Z_i) = \sum_{i=1}^{N} \pi_i(1-\pi_i).$$

(b) Because the random variables Z_i are independent,

$$\begin{aligned} P\left(\sum_{i=1}^{N} Z_i = 0\right) &= P\left(Z_1 = 0, Z_2 = 0, \ldots, Z_N = 0\right) \\ &= \prod_{i=1}^{N} P(Z_i = 0) \\ &= \prod_{i=1}^{N} (1-\pi_i). \end{aligned}$$

(c) Recall that $e^x = \lim_{N\to\infty}\left(1+\frac{x}{N}\right)^N$. Suppose $\pi_i = n/N$. Then

$$P\left(\sum_{i=1}^{N} Z_i = 0\right) = \prod_{i=1}^{N}\left(1-\frac{n}{N}\right) = \left(1-\frac{n}{N}\right)^N \approx e^{-n}.$$

This probability goes to 0 as $n \to \infty$.

(d)

$$E[\hat{t}_{HT}] = \sum_{i=1}^{N} \frac{y_i}{\pi_i} E[Z_i] = \sum_{i=1}^{N} \frac{y_i}{\pi_i} E[\pi_i] = t$$

(e) Because the random variables $\{Z_1, \ldots, Z_N\}$ are independent, $\pi_{ij} = P(Z_i = 1, Z_j = 1) = \pi_i\pi_j$. This comes from the second term of (6.26) or can be proven directly. Then

$$\begin{aligned} V(\hat{t}_{HT}) &= V\left(\sum_{i=1}^{N} Z_i \frac{y_i}{\pi_i}\right) \\ &= E\left[\sum_{i=1}^{N}\sum_{k=1}^{N} Z_i Z_k \frac{y_i}{\pi_i}\frac{y_k}{\pi_k}\right] - t^2 \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + \sum_{i=1}^{N} \sum_{k \neq i} y_i y_k - t^2 \\
&= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + \sum_{i=1}^{N} \sum_{k=1}^{N} y_i y_k - \sum_{i=1}^{N} y_i^2 - t^2 \\
&= \sum_{i=1}^{N} y_i^2 \left(\frac{1}{\pi_i} - 1 \right) \\
&= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} (1 - \pi_i).
\end{aligned}
$$

(f) This follows directly from Theorem 6.4.

6.29 (a) We have π_i's 0.50 0.25 0.50 0.75, $V(\hat{t}_{\mathrm{HT}}) = 180.1147$, and $V(\hat{t}_{\psi}) = 101.4167$.

(b)

$$
\begin{aligned}
&\frac{1}{2n} \sum_{i=1}^{N} \sum_{k=1}^{N} \pi_i \pi_k \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 \\
&= \frac{1}{2n} \sum_{i=1}^{N} \sum_{k=1}^{N} \pi_i \pi_k \left[\left(\frac{t_i}{\pi_i} - \frac{t}{n} \right) - \left(\frac{t_k}{\pi_k} - \frac{t}{n} \right) \right]^2 \\
&= \frac{1}{2n} \sum_{i=1}^{N} \sum_{k=1}^{N} \pi_i \pi_k \left[\left(\frac{t_i}{\pi_i} - \frac{t}{n} \right)^2 + \left(\frac{t_k}{\pi_k} - \frac{t}{n} \right)^2 - 2 \left(\frac{t_i}{\pi_i} - \frac{t}{n} \right) \left(\frac{t_k}{\pi_k} - \frac{t}{n} \right) \right] \\
&= \frac{1}{n} \sum_{i=1}^{N} \sum_{k=1}^{N} \pi_i \pi_k \left(\frac{t_i}{\pi_i} - \frac{t}{n} \right)^2 - \frac{1}{n} \left[\sum_{i=1}^{N} \pi_i \left(\frac{t_i}{\pi_i} - \frac{t}{n} \right) \right]^2 \\
&= \frac{1}{n} \sum_{i=1}^{N} \sum_{k=1}^{N} n^2 \psi_i \psi_k \left(\frac{t_i}{n \psi_i} - \frac{t}{n} \right)^2 - \frac{1}{n} \left[\sum_{i=1}^{N} (t_i - \psi_i t) \right]^2 \\
&= \frac{1}{n} \sum_{i=1}^{N} \psi_i \left(\frac{t_i}{\psi_i} - t \right)^2 \\
&= V(\hat{t}_{\psi}).
\end{aligned}
$$

(c)

$$
\begin{aligned}
V(\hat{t}_{\psi}) - V(\hat{t}_{HT}) &= \frac{1}{2n} \sum_{i=1}^{N} \sum_{k=1}^{N} \pi_i \pi_k \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 \\
&\quad - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\left(\frac{\pi_i\pi_k}{n}-\pi_i\pi_k+\pi_{ik}\right)\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&> \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\left[\pi_i\pi_k\left(\frac{1}{n}-1\right)-\frac{n-1}{n}\pi_i\pi_k\right]\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2 \\
&= 0.
\end{aligned}
$$

(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,

$$\frac{n-1}{n}\sum_{i=1}^{N}\sum_{k=1}^{N}t_it_k \geq \sum_{i=1}^{N}\sum_{k\neq i}^{N}\pi_{ik}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k},$$

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 ≥ 0. Using a 2×2 submatrix, we have $1 - a_{ik}^2 \geq 0$, which gives the result.

6.30 (a) We have

		k				
	π_{ik}	1	2	3	4	π_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
	π_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.

6.31 (a)

$$\begin{aligned}
&P\{\text{psu's } i \text{ and } j \text{ are in the sample}\} \\
&\quad = P\{\text{psu } i \text{ drawn first and psu } j \text{ drawn second}\} \\
&\qquad\quad + P\{\text{psu } j \text{ drawn first and psu } i \text{ drawn second}\} \\
&\quad = \frac{a_i}{\sum_{k=1}^{N} a_k}\frac{\psi_j}{1-\psi_i} + \frac{a_j}{\sum_{k=1}^{N} a_k}\frac{\psi_i}{1-\psi_j} \\
&\quad = \frac{\psi_i(1-\psi_i)\psi_j}{\sum_{k=1}^{N} a_k(1-\psi_i)(1-\pi_i)} + \frac{\psi_j(1-\psi_j)\psi_i}{\sum_{k=1}^{N} a_k(1-\psi_j)(1-\pi_j)}
\end{aligned}$$

$$= \frac{\psi_i\psi_j}{\sum_{k=1}^{N} a_k}\left(\frac{1}{1-\pi_i}+\frac{1}{1-\pi_j}\right).$$

(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

$$2\sum_{k=1}^{N} a_k = 2\sum_{k=1}^{N}\frac{\psi_k(1-\psi_k)}{1-2\psi_k} = \sum_{k=1}^{N}\frac{\psi_k(1-2\psi_k+1)}{1-2\psi_k} = 1+\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}.$$

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),

$$4\left(\sum_{k=1}^{N} a_k\right)(1-\pi_i)(1-\pi_j)-(1-\pi_j+1-\pi_i)$$

$$
\begin{aligned}
&= 2\Big[1+\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}\Big](1-\pi_i)(1-\pi_j)-2+\pi_i+\pi_j\\
&= 2(1-\pi_i)(1-\pi_j)\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}+2-2\pi_i-2\pi_j+2\pi_i\pi_j-2+\pi_i+\pi_j\\
&= 2(1-\pi_i)(1-\pi_j)\sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}-\pi_i-\pi_j+2\pi_i\pi_j\\
&\geq 2(1-\pi_i)\psi_j+2(1-\pi_j)\psi_i-\pi_i-\pi_j+2\pi_i\pi_j\\
&= (1-\pi_i)\pi_j+(1-\pi_j)\pi_i-\pi_i-\pi_j+2\pi_i\pi_j\\
&= 0.
\end{aligned}
$$

Thus $\pi_i\pi_j-\pi_{ij}\geq 0$, and the SYG estimator of the variance is guaranteed to be nonnegative.

6.32 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	π_i	ψ_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 π_{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	π_{ij}	π_i	π_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

$$\Sigma P(\mathcal{S})\hat{t}_{HT\mathcal{S}} = 128 \quad \text{and} \quad \Sigma P(\mathcal{S})(\hat{t}_{HT\mathcal{S}} - 128)^2 = 243.07.$$

6.33 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 lth 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,

$$P(\text{reject first } l \text{ SRSWR's}) = \Big(\sum_{k=1}^{N} \psi_k^2\Big)^l.$$

Thus

$$\begin{aligned}
\pi_{ij} &= P\{\text{psu's } i \text{ and } j \text{ are in the sample}\} \\
&= \sum_{l=0}^{\infty} P\{\text{psu's } i \text{ and } j \text{ chosen in } (l+1)\text{st SRSWR,} \\
&\qquad\qquad \text{and the first } l \text{ SRSWR's are rejected}\} \\
&= \sum_{l=0}^{\infty} (2\psi_i\psi_j)\Big(\sum_{k=1}^{N} \psi_k^2\Big)^l \\
&= \frac{2\psi_i\psi_j}{1-\sum_{k=1}^{N}\psi_k^2}.
\end{aligned}$$

Equation (6.18) implies

$$\begin{aligned}
\pi_i &= \sum_{j=1, j\neq i}^{N} \pi_{ij} \\
&= \sum_{j=1, j\neq i}^{N} 2\psi_i\psi_j \Big/ \Big(1-\sum_{k=1}^{N}\psi_k^2\Big) \\
&= 2\psi_i \Big/ \Big(1-\sum_{k=1}^{N}\psi_k^2\Big) - 2\psi_i^2 \Big/ \Big(1-\sum_{k=1}^{N}\psi_k^2\Big) \\
&= 2\psi_i(1-\psi_i) \Big/ \Big(1-\sum_{k=1}^{N}\psi_k^2\Big).
\end{aligned}$$

Note that, as (6.17) predicts,

$$\sum_{i=1}^{N} \pi_i = 2\Big(1-\sum_{i=1}^{N}\psi_i^2\Big) \Big/ \Big(1-\sum_{k=1}^{N}\psi_k^2\Big) = 2$$

6.34 *Rao-Hartley-Cochran.* 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

$$\hat{t}_{\mathrm{RHC}} = \sum_{k=1}^{n} \frac{t_{\alpha(k)}}{x_{k,\alpha(k)}} = \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} Z_i \frac{t_i}{x_{ki}}.$$

We show that $\hat{t}_{\mathrm{RHC}}$ is conditionally unbiased for t given the grouping

$$\begin{aligned} E[\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN}] &= E\Big[\sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} Z_i \frac{t_i}{x_{ki}} \mid I_{11}, \ldots, I_{nN} \Big] \\ &= \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} \frac{t_i}{x_{ki}} \sum_{l=1}^{N} I_{kl} x_{kl} \\ &= \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} t_i \\ &= \sum_{i=1}^{N} t_i = t. \end{aligned}$$

Since $E[\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN}] = t$ for any random grouping of psu's, we have that $E[\hat{t}_{\mathrm{RHC}}] = t$.

To find the variance, note that

$$V[\hat{t}_{\mathrm{RHC}}] = E[V(\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN})] + V[E(\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN})].$$

Since $E[\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN}] = t$, however, we know that $V[E(\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN})] = 0$. Conditionally on the grouping, the kth term in $\hat{t}_{\mathrm{RHC}}$ 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

$$\begin{aligned} V(\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN}) &= \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} x_{ki} \Big(\frac{t_i}{x_{ki}} - \sum_{j=1}^{N} I_{kj} t_j \Big)^2 \\ &= \sum_{k=1}^{n}\sum_{i=1}^{N} I_{ki} \frac{t_i^2}{x_{ki}} - \sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{N} I_{ki} t_i I_{kj} t_j \\ &= \sum_{k=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N} I_{ki} I_{kj} \Big(\frac{M_j t_i^2}{M_i} - t_i t_j \Big). \end{aligned}$$

item Now to find $E[V(\hat{t}_{\mathrm{RHC}} \mid I_{11}, \ldots, I_{nN})]$, 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 $\psi_i = M_i / \sum_{j=1}^N M_j$,

$$\begin{aligned}
V[\hat{t}_{\text{RHC}}] &= E[V(\hat{t}_{\text{RHC}} \mid I_{11}, \ldots, I_{nN})] \\
&= E\Big[\sum_{k=1}^n \sum_{i=1}^N \sum_{j=1}^N I_{ki} I_{kj} \Big(\frac{M_j t_i^2}{M_i} - t_i t_j\Big)\Big] \\
&= \sum_{k=1}^n \sum_{i=1}^N \frac{N_k}{N}(t_i^2 - t_i^2) + \sum_{k=1}^n \sum_{i=1}^N \sum_{j=1, j\neq i}^N \frac{N_k}{N}\frac{N_k - 1}{N-1}\Big(\frac{M_j t_i^2}{M_i} - t_i t_j\Big) \\
&= \Big(\sum_{k=1}^n \frac{N_k}{N}\frac{N_k-1}{N-1}\Big)\Big(\sum_{i=1}^N \frac{t_i^2}{\psi_i} - t^2\Big) \\
&= \Big(\sum_{k=1}^n \frac{N_k}{N}\frac{N_k-1}{N-1}\Big)\Big(\sum_{i=1}^N \psi_i \Big[\frac{t_i}{\psi_i} - t\Big]^2\Big).
\end{aligned}$$

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,

$$\begin{aligned}
V[\hat{t}_{\text{RHC}}] &= \frac{L-1}{N-1}\sum_{i=1}^N \psi_i \Big(\frac{t_i}{\psi_i} - t\Big)^2 \\
&= \frac{N-n}{N-1} V(\hat{t}_\psi).
\end{aligned}$$

6.35 (a) They do not satisfy condition (6.18).

$$\begin{aligned}
\sum_{k\neq i} \tilde{\pi}_{ik} &= \sum_{k\neq i} \pi_i \pi_k \left[1 - (1-\pi_i)(1-\pi_k)/\sum_{j=1}^N c_j\right] \\
&= \pi_i \sum_{k\neq i} \pi_k - \frac{\pi_i(1-\pi_i)}{\sum_{j=1}^N c_j} \sum_{k\neq i} \pi_k(1-\pi_k) \\
&= \pi_i(n - \pi_i) - \frac{\pi_i(1-\pi_i)}{\sum_{j=1}^N c_j}\left[\sum_{k=1}^N \pi_k(1-\pi_k) - \pi_i(1-\pi_i)\right] \\
&= \pi_i(n-\pi_i) - \pi_i(1-\pi_i) + \frac{\pi_i^2(1-\pi_i)^2}{\sum_{j=1}^N c_j} \\
&= \pi_i(n-1) + \frac{\pi_i^2(1-\pi_i)^2}{\sum_{j=1}^N c_j}.
\end{aligned}$$

(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 $B = \sum_{j=1}^{N} c_j$,

$$\begin{aligned}
V_{Haj}(\hat{t}_{HT}) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_i\pi_k - \tilde{\pi}_{ik})\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k\frac{(1-\pi_i)(1-\pi_k)}{\sum_{j=1}^{N} c_j}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 \\
&= \frac{1}{2\sum_{j=1}^{N} c_j}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k(1-\pi_i)(1-\pi_k)\left(2\frac{t_i^2}{\pi_i^2} - 2\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right) \\
&= \frac{1}{\sum_{j=1}^{N} c_j}\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k(1-\pi_i)(1-\pi_k)\left(\frac{t_i^2}{\pi_i^2} - \frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right) \\
&= \sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{t_i^2}{\pi_i^2} - \frac{1}{\sum_{j=1}^{N} c_j}\left(\sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{t_i}{\pi_i}\right)^2 \\
&= \sum_{i=1}^{N} c_i\frac{t_i^2}{\pi_i^2} - \frac{1}{\sum_{j=1}^{N} c_j}\left(\sum_{i=1}^{N} c_i\frac{t_i}{\pi_i}\right)^2
\end{aligned}$$

$$= \sum_{i=1}^{N} c_i \left(\frac{t_i^2}{\pi_i^2} - A \right)^2 .$$

6.36 (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

$$\sum_{i=1}^{N} \frac{n}{N} \left(\frac{N t_i}{n} - \frac{t}{n} \right)^2 = \sum_{i=1}^{N} \frac{N}{n} (t_i - \bar{t}_U)^2 = N(N-1) \frac{S_t^2}{n}.$$

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 π_{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

$$\sum_{i=1}^{N}\pi_i(1-c_i\pi_i)\left(\frac{t_i}{\pi_i} - \frac{t}{n}\right)^2 \;=\; \frac{N(N-1)}{n}\left(1 - \frac{n(n-1)}{N-n}\right)S_t^2.$$

6.37 We wish to minimize

$$\frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2S_i^2}{m_i\psi_i}$$

subject to the constraint that

$$C = E\Big[\sum_{i\in\mathcal{S}} m_i\Big] = E\Big[\sum_{i=1}^{N} Q_i m_i\Big] = n\sum_{i=1}^{N}\psi_i m_i.$$

Using Lagrange multipliers, let

$$g(m_1,\ldots,m_N,\lambda) = \sum_{i=1}^{N}\frac{M_i^2S_i^2}{m_i\psi_i} - \lambda\Big(C - n\sum_{i=1}^{N}\psi_i m_i\Big).$$

$$\frac{\partial g}{\partial m_k} = -\frac{M_k^2S_k^2}{m_k^2\psi_k} + n\lambda\psi_k$$
$$\frac{\partial g}{\partial \lambda} = n\sum_{i=1}^{N}\psi_i m_i - C.$$

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 $m_i \propto M_i/\psi_i$.

6.38 *Mitofsky-Waksberg method.* Let M_i be the number of residential numbers in psu i. When you dial a number based on the method,

$$P(\text{reach working number in psu } i \text{ on an attempt})$$

$$
\begin{aligned}
&= P(\text{select psu } i)P(\text{get working number} \mid \text{select psu } i) \\
&= \frac{1}{N}\frac{M_i}{100}.
\end{aligned}
$$

Also,

$$
P(\text{reach no number on an attempt}) = \sum_{i=1}^{N} \frac{1}{N}\left(1 - \frac{M_i}{100}\right) = 1 - \frac{M_0}{100N}.
$$

Then,

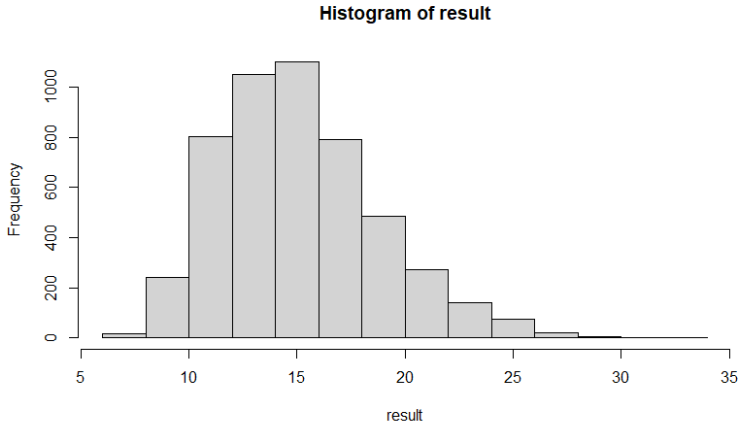
$$
\begin{aligned}
&P(\text{select psu } i \text{ as first in sample}) \\
&= P(\text{select psu } i \text{ on first attempt}) \\
&\quad + P(\text{reach no one on first attempt, select psu } i \text{ on second attempt}) \\
&\quad + P(\text{reach no one on first and second attempts,} \\
&\qquad \text{select psu } i \text{ on third attempt}) + \ldots \\
&= \frac{1}{N}\frac{M_i}{100} + \frac{1}{N}\frac{M_i}{100}\left(1 - \frac{M_0}{100N}\right) + \frac{1}{N}\frac{M_i}{100}\left(1 - \frac{M_0}{100N}\right)^2 + \ldots \\
&= \frac{1}{N}\frac{M_i}{100}\sum_{j=0}^{\infty}\left(1 - \frac{M_0}{100N}\right)^j \\
&= \frac{1}{N}\frac{M_i}{100}\frac{1}{1 - \left(1 - \frac{M_0}{100N}\right)} \\
&= \frac{M_i}{M_0}
\end{aligned}
$$

6.45 I used the following R code to do the simulation:

```
sumwt <- function(prob,n,numsamp=1000) {
 result <- rep(0,numsamp)
 for (i in 1:numsamp) {
   sample_units<-sample((1:length(prob)),n,replace=TRUE,prob=prob)
# calculate ExpectedHits and sampling weights
   ExpectedHits<-n*prob[sample_units]/sum(prob)
   SamplingWeight<-1/ExpectedHits
# check sum of sampling weights
  result[i]=sum(SamplingWeight)
}
result
}
result = sumwt(prob=classes$class_size,5,5000)
summary(result)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.47   12.32   14.67   15.08   17.26   32.06

sqrt(var(result))
```

```
[1] 3.658185
hist(result)
```



Chapter 7

Complex Surveys

7.1–**7.6** Answers will vary.

7.7 Here is SAS software code for solving this problem . Note that for the population, we have $\bar{y}_{\mathcal{U}} = 17.73$,

$$S^2 = 38.1451726 = \frac{1}{1999}\left[\sum_{i=1}^{2000} y_i^2 - \frac{1}{2000}\left(\sum_{i=1}^{2000} y_i\right)^2\right] = \left(704958 - \frac{35460^2}{2000}\right)/1999,$$

$\hat{\theta}_{25} = 13.098684, \hat{\theta}_{50} = 16.302326$, and $\hat{\theta}_{75} = 19.847458$.

```
data integerwt;
   set integerwt;
   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).

7.8 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;
   set datalib.nybight;
   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;

/* The FREQ procedure gives values in percents, so divide each by 100*/
data htpop_epmf;
    set htpop_epmf;
        epmf = percent/100;
        ecdf =  cum_pct/100;

proc print data=htpop_epmf;
run;
```

7.9 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. (You could alternatively take m_i to be the number of forms.) 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

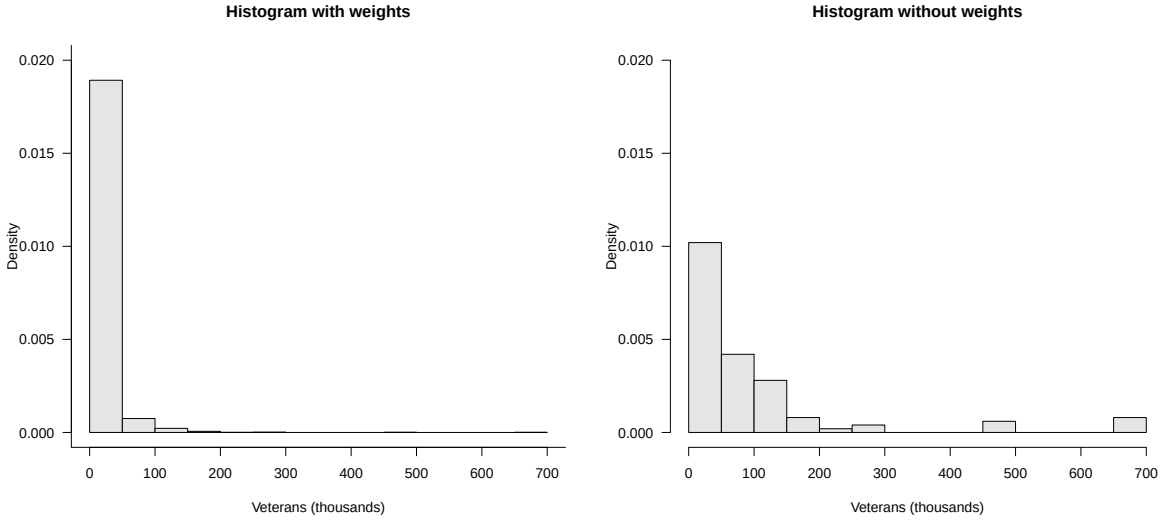
y	$\hat{f}(y)$	y	$\hat{f}(y)$
27.50	0.0192	35.85	0.0164
27.90	0.0125	36.20	0.0022
28.30	0.0050	36.25	0.0421
29.15	0.0177	36.65	0.0664
30.00	0.0375	37.05	0.0023
30.40	0.0359	37.10	0.0403
30.80	0.0031	37.50	0.1307
31.25	0.0662	37.90	0.0079
32.05	0.0022	37.95	0.0019
32.10	0.0031	38.35	0.0163
32.50	0.0370	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 The design effect for returning a form is 0.98; the deff for having had measles is 6.1.

7.11 The deff for *mathlevel* is 2.85.

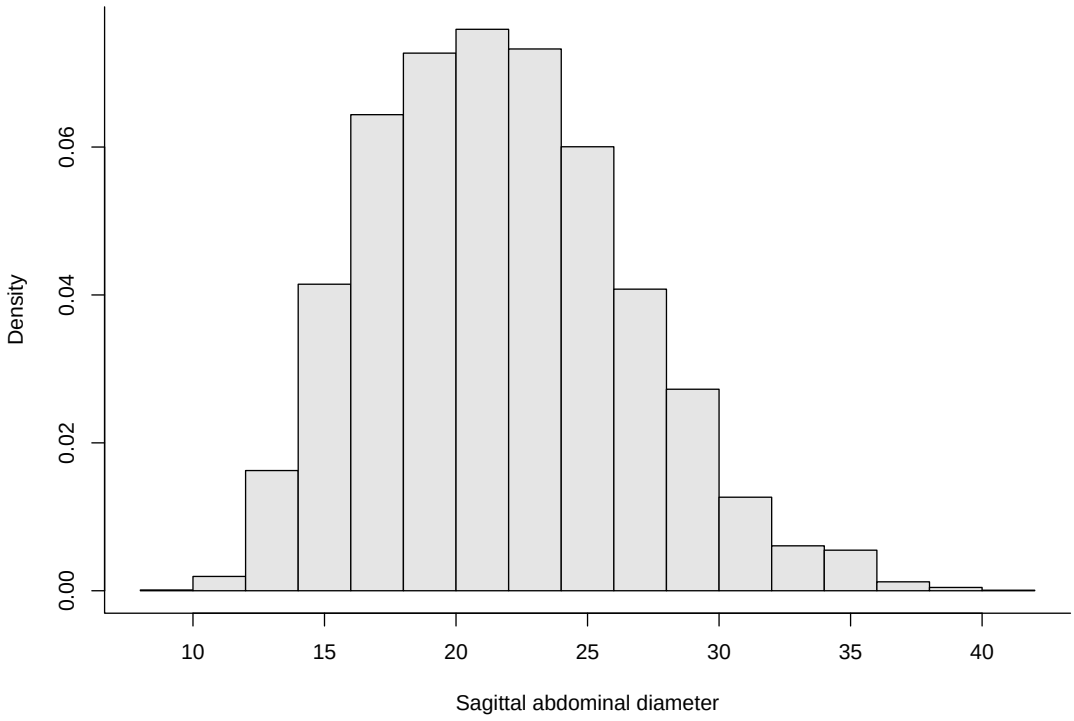
7.12 The two histograms are shown in Figure SM7.1.

Figure 7.1: Histograms of veterans, with and without weights, for Exercise 7.12.



7.14 (a) Figure SM7.2 shows the histogram.

Figure 7.2: Histogram for Exercise 7.14.



(b) Here is output from R.

```
data(nhanes)
nhanes$age20d<-rep(0,nrow(nhanes))
nhanes$age20d[nhanes$ridageyr >= 20 & !is.na(nhanes$bmdavsad)]<- 1
dnhanes<-svydesign(id=~sdmvpsu,strata=~sdmvstra,weights=~wtmec2yr,nest=TRUE, data=nhanes)
dsub<-subset(dnhanes,age20d==1)
sadmean<-svymean(~bmdavsad,dsub)
sadmean
##             mean     SE
## bmdavsad 22.781 0.1957

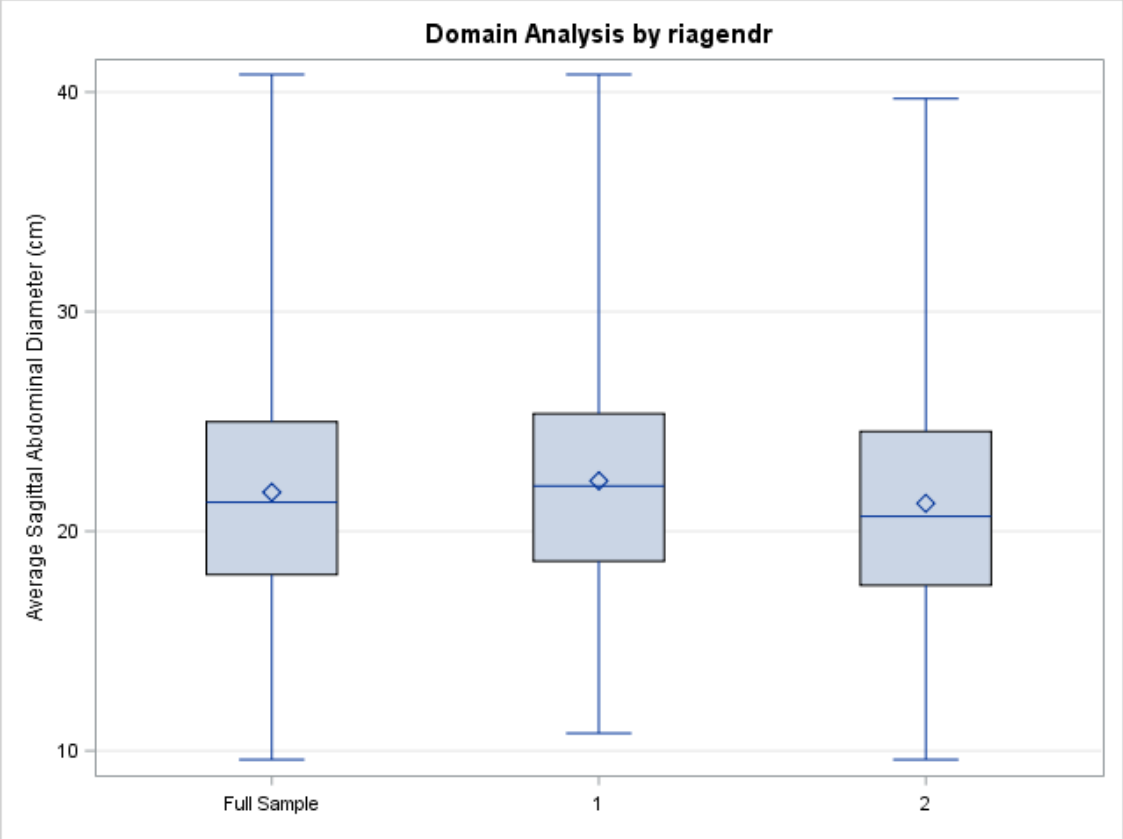
confint(sadmean,df=15)
##              2.5 %   97.5 %
## bmdavsad 22.36413 23.19839
```

(c) Here are the quantiles:

Quantile	0	0.25	0.5	0.75	1
All	9.6	18.02048	21.32016	24.9836	40.8
Male	10.800000	18.634896	22.061504	25.352110	40.800000
Female	9.600000	17.533723	20.678681	24.546721	39.700000

Figure SM7.3 shows the boxplot.

Figure 7.3: Boxplot for Exercise 7.14 (1 = male, 2 = female).



7.15 (b)

Quantile	0	0.25	0.5	0.75	1
All	40.000000	79.471878	93.986937	106.737584	171.600000
Male	40.000000	80.751256	96.100379	108.007754	169.600000
Female	40.800000	78.668843	91.914578	105.539945	171.600000

7.16 Using the formula for stratified sampling and ignoring the fpc,

$$V\left(\bar{y}_{str,d}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h} = \sum_{h=1}^{H} \frac{N_h}{N} \frac{N_h}{N n_h} S_1^2 = \sum_{h=1}^{H} \frac{N_h}{N} \frac{w_h}{N} S_1^2.$$

Similarly,

$$V\left(\bar{y}_{str,p}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{(nN_h/N)} = \sum_{h=1}^{H} \frac{N_h}{N} \frac{S_1^2}{n} = \frac{S_1^2}{n}.$$

Thus,

$$\frac{V\left(\bar{y}_{str,d}\right)}{V\left(\bar{y}_{str,p}\right)} = \sum_{h=1}^{H} \frac{N_h}{N} w_h \frac{n}{N} = \left(\sum_{h=1}^{H} \frac{N_h}{N} w_h\right) \left(\sum_{h=1}^{H} \frac{N_h}{N w_h}\right).$$

The last step is true because

$$\sum_{h=1}^{H} \frac{N_h}{N w_h} = \sum_{h=1}^{H} \frac{N_h n_h}{N N_h} = \frac{n}{N}.$$

7.17 We derive the weighting effect (sometimes called the weff) under a superpopulation model. Suppose that $Y_1, \ldots, Y_n$ are iid with mean μ and variance σ^2. Then

$$V\left(\frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}\right) = \frac{\sum_{i=1}^n w_i^2 \sigma^2}{\left(\sum_{i=1}^n w_i\right)^2}$$

The variance of the unweighted mean is simply σ^2/n. Thus the design effect due to weighting is

$$\begin{aligned}
\text{weff} &= \frac{\left(\sum_{i=1}^n w_i^2 \sigma^2\right) / \left(\sum_{i=1}^n w_i\right)^2}{\sigma^2/n} \\
&= n \frac{\sum_{i=1}^n w_i^2}{\left(\sum_{i=1}^n w_i\right)^2} \\
&= 1 + n \frac{\sum_{i=1}^n w_i^2}{\left(\sum_{i=1}^n w_i\right)^2} - \frac{\left(\sum_{i=1}^n w_i\right)^2}{\left(\sum_{i=1}^n w_i\right)^2} \\
&= 1 + \frac{n \sum_{i=1}^n w_i^2 - \left(\sum_{i=1}^n w_i\right)^2}{\left(\sum_{i=1}^n w_i\right)^2} \\
&= 1 + \frac{n \left(\sum_{i=1}^n w_i^2 - n\bar{w}^2\right)}{n^2 \bar{w}^2} \\
&= 1 + \frac{n-1}{n} \frac{s_w^2}{\bar{w}^2},
\end{aligned}$$

which is approximately 1 plus the squared coefficient of variation of the weights.

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.

7.19 (a) First note that because $y_{(1)} < y_{(2)} < \ldots < y_{(K)}$ are the distinct values taken on by observations in the sample, $\hat{F}(y_{(k+1)}) - \hat{F}(y_{(k)}) = \hat{f}(y_{(k+1)})$. For $\hat{F}(y_{(k)}) \le q < \hat{F}(y_{(k+1)})$, using the definition of $\tilde{F}$ in (7.13),

$$\begin{aligned}
\tilde{F}(\tilde{\theta}_q) &= \tilde{F}\left[y_{(k)} + \frac{q - \hat{F}(y_{(k)})}{\hat{F}(y_{(k+1)}) - \hat{F}(y_{(k)})}(y_{(k+1)} - y_{(k)})\right] \\
&= \hat{F}\left[y_{(k)}\right] + \frac{\tilde{\theta}_q - y_{(k)}}{y_{(k+1)} - y_{(k)}}\hat{f}(y_{(k+1)}) \\
&= \hat{F}\left[y_{(k)}\right] + \frac{q - \hat{F}(y_{(k)})}{\hat{F}(y_{(k+1)} - \hat{F}(y_{(k)})}(y_{(k+1)} - y_{(k)})\frac{\hat{f}(y_{(k+1)})}{y_{(k+1)} - y_{(k)}} \\
&= \hat{F}(y_{(k)}) + q - \hat{F}(y_{(k)}) \\
&= q.
\end{aligned}$$

(b) Figure SM7.4 shows the interpolated empirical cdf $\tilde{F}$ along with the step function of $\hat{F}$. The interpolated cdf connects the left ends of the lines of the empirical cdf.

(c) Relevant quantities for the calculations are:

q	$y_{(k)}$	$\hat{F}(y_{(k)})$	$y_{(k+1)}$	$\hat{F}(y_{(k+1)})$	$y_{(k)} + \frac{q - \hat{F}(y_{(k)})}{\hat{F}(y_{(k+1)} - \hat{F}(y_{(k)})}(y_{(k+1)} - y_{(k)})$
0.25	160	0.234375	161	0.25625	160.7142857
0.5	167	0.484375	168	0.5125	167.5555556
0.75	176	0.734375	177	0.759375	176.625
0.9	181	0.8875	182	0.9125	181.5

These are the quantiles calculated by the SURVEYMEANS procedure. Here is the code and output:

```
proc surveymeans data=htstrat mean sum percentile=(25 50 75 90);
   weight wt;
   strata gender;
   var height;
run;
```

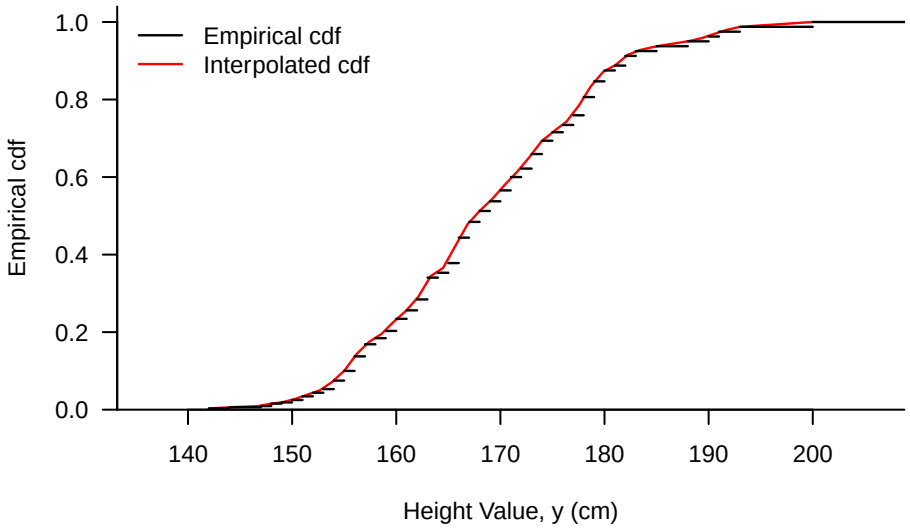


Figure 7.4: Interpolated empirical cdf, and empirical cdf, for Exercise 7.19.

Quantiles						
Variable	Percentile		Estimate	Std Error	95% Confidence Limits	
height	25	Q1	160.714286	0.766758	159.202226	162.226346
	50	Median	167.555556	1.042272	165.500177	169.610934
	75	Q3	176.625000	1.353653	173.955572	179.294428
	90	D9	181.500000	2.579060	176.414049	186.585951

7.22 As stated in Section 7.1, the y_is 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

$$\sum_y y\hat{f}(y) = \frac{\sum_{i \in \mathcal{S}} y_i w_i}{\sum_{i \in \mathcal{S}} w_i}$$

$$
\begin{aligned}
&= \frac{\sum_{i\in\mathcal{S}} y_i \sum_{h=1}^{H} (N_h/n_h) x_{hi}}{\sum_{i\in\mathcal{S}} \sum_{h=1}^{H} (N_h/n_h) x_{hi}} \\
&= \frac{\sum_{h=1}^{H} N_h \sum_{i\in\mathcal{S}} (x_{hi} y_i/n_h)}{\sum_{h=1}^{H} N_h \sum_{i\in\mathcal{S}} (x_{hi}/n_h)} \\
&= \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h} \\
&= \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h.
\end{aligned}
$$

7.23 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).

7.24 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 $P(D_i = 1 \mid Z_i = 1) \propto 1/\pi_i$.

$$\begin{aligned} P(i \in \mathcal{S}_2) &= P(Z_i = 1, D_i = 1) \\ &= P(D_i = 1 \mid Z_i = 1) P(Z_i = 1) \\ &\propto \frac{1}{\pi_i} \pi_i = 1. \end{aligned}$$

7.25 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.

7.26 (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 π 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×(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

7.28 (a) If the sample really were a stratified random sample, the variance of $\bar{y}$ from the systematic sample is estimated by

$$\hat{V}_{IS}(\bar{y}) = \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{(N/n)^2}{N^2} \frac{s_h^2}{2} = \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{s_h^2}{2n^2},$$

where s_h^2 is the sample variance of the two units assumed to come from stratum h.

Chapter 8

Nonresponse

8.1 (a) Using a voter file reduced calls to non-households and non-voters that would be made under random digit dialing. However, persons who registered to vote after the file was assembled, persons without a telephone number in the database, and persons with an incorrect telephone number were not contacted.

(b) Not necessarily. They could agree on everything except the variable of interest.

(c) If information is available on other variables for the registered voters, one might see how these variables were associated with turnout and outcomes in different districts, and adopt some of the additional variables for weighting.

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,

$$\bar{y} = \frac{5894}{150} = 39.3$$
$$s^2 = \frac{291725 - (150)(39.3)^2}{149} = 403.6$$

and

$$\text{SE}(\bar{y}) = \sqrt{\left(1 - \frac{150}{2000}\right)\frac{403.6}{150}} = 1.58.$$

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 χ^2 distribution with 2 df, the p-value is 1.4×10^{-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)

$$\text{SSB} = \sum_{i=1}^{3} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = 9303.1$$

$$\text{MSW} = s^2 = 403.6.$$

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 ≥ 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 ≥ 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_{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
$$\hat{y}_{\text{post}} = \frac{82140}{2000} = 41.07.$$

8.5 (a) Response rates are given in Table SM8.1:

(b) Yes! (If they weren't would I have put this exercise in the textbook?) The response rates differ greatly across the four classes, indicating that the family head's race and gender are associated

Table 8.1: Response rates and weighting for Exercise 8.5

Family Head's Gender	Family Head's Race	Sample Size	Number of Respondents	Response Rate (%)	Nonresponse-adjusted Weight
Male	White	160	77	48.1	103.90
Male	Nonwhite	51	35	68.6	72.86
Female	White	251	209	83.3	60.05
Female	Nonwhite	358	327	91.3	54.74
Total		820	648	79.0	

with the response propensities. This means that using these categories for weighting is likely to reduce the bias for any y variable measured in the survey.

The mean income of the respondents vary across the classes too, so the weighting will reduce the bias for this particular response, and probability also reduce the variance (if the effect of increasing the homogeneity within classes is greater than the effect of having unequal weighting).

(c) Table SM8.1 gives the nonresponse-adjusted weights. Each is calculated as 50 × (sample size) / (number of respondents). The weight adjustments are larger for classes with lower response rates.

(d) With the sampling weights, the average income is

$$\frac{50 \times (77 * 1712 + 35 * 1509 + 209 * 982 + 327 * 911)}{50 \times 648} = \$1{,}061.38.$$

With the nonresponse-adjusted weights, the average income is

$$\frac{77 * 103.90 * 1712 + 35 * 72.86 * 1509 + 209 * 60.05 * 982 + 327 * 54.74 * 911}{77 * 103.90 + 35 * 72.86 + 209 * 60.05 + 327 * 54.74} = \$1{,}126.22.$$

The average income with the nonresponse-adjusted weights is higher because the groups with higher incomes have lower response rates.

8.6 First, we need to find the sum of the nonresponse-adjusted weights for each poststratum. Because these differ by the race and gender of the family head, we set up a spreadsheet for the 16 gender × race × education × employment categories. Note that in the last column I check that the weights sum to 50 × 820, the sum of the sampling weights for the full sample.

Gender	Race	HS Diploma?	Employed?	Respondents	NR-adj. Weight	Sum of NR-adj Weights
Male	White	No	No	16	103.90	1662.34
Male	Nonwhite	No	No	9	72.86	655.71
Female	White	No	No	45	60.05	2702.15
Female	Nonwhite	No	No	71	54.74	3886.54
Male	White	Yes	No	12	103.90	1246.75
Male	Nonwhite	Yes	No	5	72.86	364.29
Female	White	Yes	No	33	60.05	1981.58
Female	Nonwhite	Yes	No	51	54.74	2791.74
Male	White	No	Yes	11	103.90	1142.86
Male	Nonwhite	No	Yes	6	72.86	437.14
Female	White	No	Yes	30	60.05	1801.44
Female	Nonwhite	No	Yes	47	54.74	2572.78
Male	White	Yes	Yes	38	103.90	3948.05
Male	Nonwhite	Yes	Yes	15	72.86	1092.86
Female	White	Yes	Yes	101	60.05	6064.83
Female	Nonwhite	Yes	Yes	158	54.74	8648.93
Total				648		41000

Now we can find the sum of the nonresponse-adjusted weights for the four poststrata:

HS Diploma?	Employed?	Sum of NR-adj Weights	Poststratum Count	Poststratification Weight Factor
No	No	8906.75	10596	1.190
Yes	No	6384.36	6966	1.091
No	Yes	5954.22	6313	1.060
Yes	Yes	19754.67	22125	1.120
Total		41000	46000	

Finally, apply the poststratification weight factor to each of the 16 classifications. Check that the sum of the weights equals the population size from the external data source, 46,000. The results are in Table SM8.2.

8.7 (a) The null hypothesis is $H_0 : p_1 = p_2 = p_3 = p_4$ where p_j is the proportion of respondents in group j. Set up the contingency table as:

	Control	Lottery	Cash	Lottery/Cash	Total
Respondents	168	198	205	281	852
Nonrespondents	832	802	795	719	3148
Total	1000	1000	1000	1000	4000

Table 8.2: Results for Exercise 8.6

Gender	Race	HS Diploma?	Employed?	Number Respond	NR-adj. Weight	Poststrat. Wt. Factor	Final Weight	Sum of Final Wts.
Male	White	No	No	16	103.90	1.190	123.60	1977.62
Male	Nonwhite	No	No	9	72.86	1.190	86.68	780.08
Female	White	No	No	45	60.05	1.190	71.44	3214.64
Female	Nonwhite	No	No	71	54.74	1.190	65.12	4623.66
Male	White	Yes	No	12	103.90	1.091	113.36	1360.34
Male	Nonwhite	Yes	No	5	72.86	1.091	79.49	397.47
Female	White	Yes	No	33	60.05	1.091	65.52	2162.11
Female	Nonwhite	Yes	No	51	54.74	1.091	59.73	3046.08
Male	White	No	Yes	11	103.90	1.060	110.16	1211.72
Male	Nonwhite	No	Yes	6	72.86	1.060	77.25	463.48
Female	White	No	Yes	30	60.05	1.060	63.67	1909.98
Female	Nonwhite	No	Yes	47	54.74	1.060	58.04	2727.81
Male	White	Yes	Yes	38	103.90	1.120	116.36	4421.77
Male	Nonwhite	Yes	Yes	15	72.86	1.120	81.60	1223.99
Female	White	Yes	Yes	101	60.05	1.120	67.25	6792.54
Female	Nonwhite	Yes	Yes	158	54.74	1.120	61.31	9686.70
Total				648				46000

The expected number of respondents in each group under the null hypothesis is (852)(1000)/4000 = 213; the expected number of nonrespondents in each group under H_0 is (3148)(1000)/4000 = 787.

The test statistic is

$$\begin{aligned} X^2 &= \frac{(168-213)^2}{213} + \frac{(198-213)^2}{213} + \frac{(205-213)^2}{213} + \frac{(281-213)^2}{213} \\ &\quad + \frac{(832-787)^2}{787} + \frac{(802-787)^2}{787} + \frac{(795-787)^2}{787} + \frac{(719-787)^2}{787} \\ &= 41.389. \end{aligned}$$

X^2 is compared to a χ^2 distribution with 3 (= $[2-1][4-1]$) df, to give p-value 5.5×10^{-9}. This is strong evidence against the null hypothesis that the response rates are equal.

(b) The null hypothesis is $H_0 : p_1 = p_2 = p_3 = p_4$ where now p_j is the proportion of respondents in group j who are male (or, if you prefer, the null hypothesis is that being male is independent of the group). The total row in the contingency table is the number of respondents. The expected count under H_0 is given in parentheses below each observed count.

	Group				
	Control	Lottery	Cash	Lottery/Cash	Total
Male, Observed	57	79	86	116	338
Male, Expected	66.64789	78.54930	81.32629	111.47653	
Female, Observed	111	119	119	165	514
Female, Expected	101.3521	119.4507	123.6737	169.5235	
Total Respondents	168	198	205	281	852

The test statistic is

$$X^2 = \sum_{\text{8cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = 3.07.$$

X^2 is compared to a χ^2 distribution with 3 ($= [2-1][4-1]$) df, to give p-value 0.38. There is no evidence that the proportions of respondents who are male differ among the groups.

There is, however, quite strong evidence that the proportion of males does not equal 0.5. For example, in the Lottery/Cash group, which has the largest sample size, the proportion of males is $\hat{p} = 116/281 = 0.413$. The test statistic

$$Z = \frac{0.413 - 0.5}{\sqrt{(0.5)(0.5)/281}} = -2.92$$

is compared to a normal distribution, giving p-value 0.004.

(c) Calculate

$$\text{SSW} = \sum_{i=1}^{4}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{4}(n_i - 1)s_i^2 = 217657.9,$$

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{4} n_i \bar{y}_i}{\sum_{i=1}^{4} n_i} = \frac{}{852} = 51.6257,$$

and

$$\text{SSB} = \sum_{i=1}^{4} n_i(\bar{y}_i - \bar{\bar{y}})^2 = 935.5271$$

to give the following ANOVA table:

Source	df	Sum of Squares	Mean Square
Between groups	3	935.5	311.8
Within groups	848	217657.9	256.7
Total, about mean	851	218593.4	

The F-statistic is $311.8/256.7 = 1.21$. Comparing this to an F distribution with 3 and 848 degrees of freedom gives p-value 0.30. There is no evidence that the mean age differs among groups.

(d) Using both incentives gives the highest response rate, and there is no evidence that it results in a different percentage of male respondents or different mean age than the other groups. However, there is still concern that the percentage of males is too small and the average age is higher than that in the community. And the response rate in group 4 of 28.1% is still pretty low.

8.8 Note that RR1 = RR3 = RR5 for this problem since there are no cases of unknown eligibility. You probably should report both response rates for this study. The unweighted response rate measures the success of the survey operations in obtaining respondents, and the weighted response rate estimates the percentage in the population who would respond.

$$RR_1 = 100 \times \frac{45}{45+17} = 72.6\%$$
$$RR_{1w} = 100 \times \frac{589.33}{589.33+262.75} = 69.2\%$$

8.9 The percentages given by U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (2014) are rounded and do not add to 100%. To get them to sum to 100%, I assumed that 31.2% were ineligible and 53.3% had unknown eligibility (note that the response rates will be similar no matter how you deal with the rounding error). With this, approximately 94,278 (= 0.467× 201,881) numbers had known eligibility and 107,603 had unknown eligibility. Estimating the eligibility rate e as (100) 31,241/94,278 = 33%, we have:

$$\text{RR1} = 100\,\frac{16{,}507}{16{,}507+1{,}542+13{,}192+107{,}603} = 11.9\%$$
$$\text{RR2} = 100\,\frac{16{,}507+1{,}542}{16{,}507+1{,}542+13{,}192+107{,}603} = 13.0\%$$
$$\text{RR3} = 100\,\frac{16{,}507}{16{,}507+1{,}542+13{,}192+(0.33)107{,}603} = 24.7\%$$
$$\text{RR4} = 100\,\frac{16{,}507+1{,}542}{16{,}507+1{,}542+13{,}192+(0.33)107{,}603} = 27.0\%$$
$$\text{RR5} = 100\,\frac{16{,}507}{16{,}507+1{,}542+13{,}192} = 52.8\%$$
$$\text{RR6} = 100\,\frac{16{,}507+1{,}542}{16{,}507+1{,}542+13{,}192} = 57.8\%$$

RR3 would be most appropriate for the report. The partial completes did not answer the questions of interest, and thus should not be counted in the response rate. RR5 and RR6 are much too high, since undoubtedly some percentage of the unanswered numbers belong to households in which someone was eligible for the survey. You might also want to report RR1 for this survey,

as it provides a conservative estimate assuming that the numbers of unknown eligibility are all nonrespondents.

Note that the response rate given by U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (2014) differed from all of these. They estimated e using "information identified by interviewers when calling numbers" and combined separate response rates for cell and landline numbers to give an overall response rate as a range from 27.5% to 33.6%.

8.11 We can select the sample and rake the counts using R. Here is sample code.

```
data(college)
# Few observations in ccsizset categories 6,7,8: combine them
# May need additional collapsing if your sample has empty cells
college$ccsizr<-college$ccsizset
college$ccsizr[college$ccsizset==6]<-8
college$ccsizr[college$ccsizset==7]<-8
set.seed(30283)
coll.srs<-getdata(college,srswor(300,nrow(college)))
head(coll.srs)
sum(is.na(coll.srs$tuitionfee_out)) # check for missing data
# check sample size in cross-classification
table(coll.srs$control,coll.srs$ccsizr)
dsrs<-svydesign(id=~1,fpc=rep(nrow(college),300),data=coll.srs)
svymean(~tuitionfee_out,dsrs)
##                 mean     SE
## tuitionfee_out 28383 588.22
unraked.quant<-svyquantile(~tuitionfee_out, dsrs, quantiles=c(0.25,0.5,0.75,0.9), ties =
    "rounded",se=TRUE,
            ci=TRUE, interval.type = "Wald")
unraked.quant$quantiles
##                 0.25   0.5  0.75   0.9
## tuitionfee_out 19360 27830 35748 43884
attributes(unraked.quant) # look at attributes
attr(unraked.quant,"SE") # extract the unraked SE
##      0.25       0.5      0.75       0.9
##  547.7051  611.7069  975.9129 1838.9989
# rake the data
# calculate population totals for the raking variables (check for missing)
ccsizr.tot <- table(college$ccsizr,useNA="ifany")
control.tot <- table(college$control,useNA="ifany")
# Create data frames containing the marginal counts
pop.ccsize <- data.frame(ccsizr=row.names(ccsizr.tot), Freq=as.vector(ccsizr.tot))
pop.control <- data.frame(control=row.names(control.tot), Freq=as.vector(control.tot))
# rake the weights
drake<- rake(dsrs, list(~ccsizr,~control), list(pop.ccsize, pop.control))
svymean(~tuitionfee_out,drake)
##                 mean     SE
## tuitionfee_out 28750 427.87
raked.quant<-svyquantile(~tuitionfee_out, drake, quantiles=c(0.25,0.5,0.75,0.9), ties = "
```

```
   rounded",se=TRUE,
            ci=TRUE, interval.type = "Wald")
raked.quant$quantiles
##                     0.25      0.5     0.75      0.9
## tuitionfee_out 19449.98 28286.49 36091.36 45043.34
attr(raked.quant,"SE")
##      0.25       0.5      0.75       0.9
##  474.1942  433.2530  772.0156 1747.4745
# look at SE ratio
attr(unraked.quant,"SE")/attr(raked.quant,"SE")
##     0.25      0.5     0.75      0.9
## 1.155023 1.411893 1.264110 1.052375
```

8.12

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 χ^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 χ^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 τ) gives p-value $> .10$. There is no evidence that the response rate is associated with the percentage of members who are female.

8.13 (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

8.15 (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.

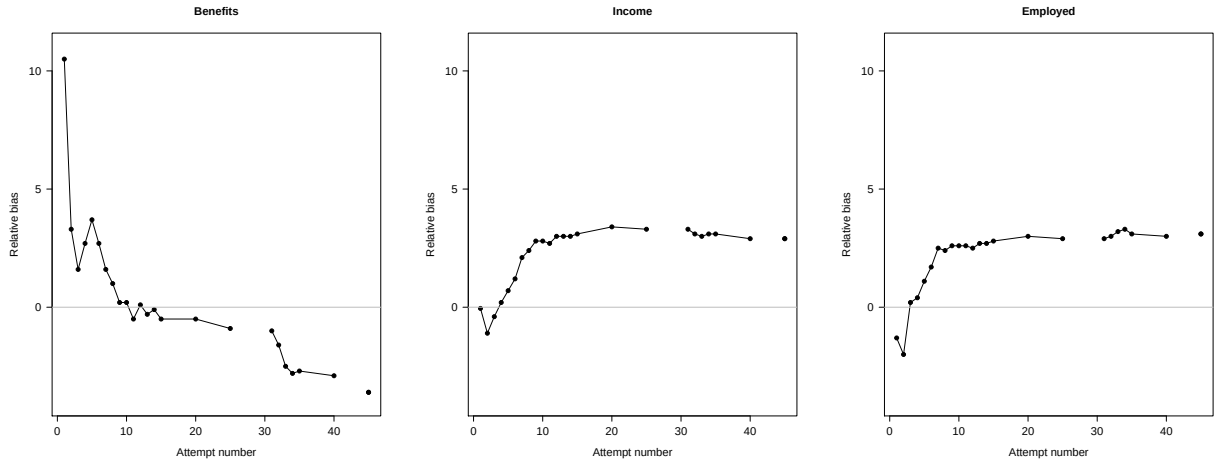
(b) The population mean of *acres92* is $\bar{y}_U = 308582$.

8.18 Figure SM8 shows the plots of the relative bias over time.

8.19 We use the approximations from Chapter 3 to obtain:

$$E[\hat{\bar{y}}_R] = E\left[\frac{\sum_{i=1}^N Z_i R_i w_i y_i}{\sum_{i=1}^N \phi_i}\left\{1 - \frac{\sum_{i=1}^N (Z_i R_i w_i - \phi_i)}{\sum_{i=1}^N Z_i R_i w_i}\right\}\right] \approx \frac{\sum_{i=1}^N \phi_i y_i}{\sum_{i=1}^N \phi_i} \approx \frac{1}{N\bar{\phi}_{\mathcal{U}}} \sum_{i=1}^N \phi_i y_i.$$

Figure 8.1: Relative bias for variables *benefits*, *income*, and *employed*



Thus the bias is

$$\begin{aligned}
\text{Bias}\,[\hat{\bar{y}}_R] &\approx \frac{1}{N\bar{\phi}_{\mathcal{U}}}\sum_{i=1}^{N}(\phi_i-\bar{\phi}_{\mathcal{U}})y_i \\
&= \frac{1}{N\bar{\phi}_{\mathcal{U}}}\sum_{i=1}^{N}(\phi_i-\bar{\phi}_{\mathcal{U}})(y_i-\bar{y}_{\mathcal{U}}) \\
&\approx \frac{1}{\bar{\phi}_{\mathcal{U}}}\text{Cov}\,(\phi_i,y_i)
\end{aligned}$$

8.20 The argument is similar to the previous exercise. We have

$$\begin{aligned}
\hat{t}_{\hat{\phi}} &= \sum_{i=1}^{N} Z_i R_i w_i y_i \sum_{c=1}^{C} \frac{x_{ic}}{\hat{\phi}_c} \\
&= \sum_{i=1}^{N} Z_i R_i w_i y_i \sum_{c=1}^{C} x_{ic} \frac{\sum_{j=1}^{N} Z_j w_j x_{jc}}{\sum_{j=1}^{N} Z_j R_i w_j x_{jc}}
\end{aligned}$$

If the weighting classes are sufficiently large, then $E[1/\hat{\phi}_c] \approx 1/\bar{\phi}_c$.

8.22 *Effect of weighting class adjustment on variances*

$$V(\hat{\bar{y}}_{\hat{\phi}}) = V\left[\frac{n_1}{n}\frac{1}{n_{1R}}\sum_{i=1}^{N} Z_i R_i x_i y_i + \frac{n_2}{n}\frac{1}{n_{2R}}\sum_{i=1}^{N} Z_i R_i (1-x_i) y_i\right]$$

$$
\begin{aligned}
&= E\left\{V\left[\frac{n_1}{n}\frac{1}{n_{1R}}\sum_{i=1}^{N} Z_i R_i x_i y_i + \frac{n_2}{n}\frac{1}{n_{2R}}\sum_{i=1}^{N} Z_i R_i (1-x_i) y_i \middle| Z_1, \ldots, Z_N\right]\right\} \\
&\quad + V\left\{E\left[\frac{n_1}{n}\frac{1}{n_{1R}}\sum_{i=1}^{N} Z_i R_i x_i y_i + \frac{n_2}{n}\frac{1}{n_{2R}}\sum_{i=1}^{N} Z_i R_i (1-x_i) y_i \middle| Z_1, \ldots, Z_N\right]\right\} \\
&= E\left\{V\left[\frac{n_1}{n}\frac{\sum_{i=1}^{N} Z_i R_i x_i y_i}{\sum_{i=1}^{N} Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^{N} Z_i R_i (1-x_i) y_i}{\sum_{i=1}^{N} Z_i R_i (1-x_i)} \middle| Z_1, \ldots, Z_N\right]\right\} \\
&\quad + V\left\{E\left[\frac{n_1}{n}\frac{\sum_{i=1}^{N} Z_i R_i x_i y_i}{\sum_{i=1}^{N} Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^{N} Z_i R_i (1-x_i) y_i}{\sum_{i=1}^{N} Z_i R_i (1-x_i)} \middle| Z_1, \ldots, Z_N\right]\right\}.
\end{aligned}
$$

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,

$$
\begin{aligned}
& V\left[\frac{n_1}{n}\frac{\sum_{i=1}^{N} Z_i R_i x_i y_i}{\sum_{i=1}^{N} Z_i R_i x_i} \middle| Z_1, \ldots, Z_N\right] \\
&= V\left[\frac{1}{n\phi_1}\sum_{i=1}^{N} Z_i R_i x_i y_i \left(1 - \frac{\sum_{i=1}^{N} Z_i [R_i - \phi_1] x_i}{\sum_{i=1}^{N} Z_i R_i x_i}\right) \middle| Z_1, \ldots, Z_N\right]
\end{aligned}
$$

$$\begin{aligned} &\approx \frac{1}{(n\phi_1)^2}\sum_{i=1}^{N} Z_i V(R_i) x_i y_i^2 \\ &\approx \frac{\phi_1(1-\phi_1)}{(n\phi_1)^2}\sum_{i=1}^{N} Z_i x_i y_i^2. \end{aligned}$$

Thus,

$$\begin{aligned} & E\left\{ V\left[\frac{n_1}{n}\frac{\sum_{i=1}^N Z_i R_i x_i y_i}{\sum_{i=1}^N Z_i R_i x_i} + \frac{n_2}{n}\frac{\sum_{i=1}^N Z_i R_i (1-x_i) y_i}{\sum_{i=1}^N Z_i R_i (1-x_i)} \Big| Z_1, \ldots, Z_N \right] \right\} \\ &\approx E\left\{ \frac{\phi_1(1-\phi_1)}{(n\phi_1)^2}\sum_{i=1}^{N} Z_i x_i y_i^2 + \frac{\phi_2(1-\phi_2)}{(n\phi_2)^2}\sum_{i=1}^{N} Z_i (1-x_i) y_i^2 \right\} \\ &= \frac{\phi_1(1-\phi_1)}{n\phi_1^2}\sum_{i=1}^{N} x_i y_i^2 + \frac{\phi_2(1-\phi_2)}{n\phi_2^2}\sum_{i=1}^{N} (1-x_i) y_i^2. \end{aligned}$$

8.23 The weighted response rate in class c is

$$\mathrm{RR}_{1w} = \frac{\sum_{i=1}^N Z_i R_i w_i x_{ic}}{\sum_{i=1}^N Z_i w_i x_{ic}},$$

where $w_i = 1/P(Z_i = 1)$. Note that the denominator is a random variable, so we have to be a bit careful about taking the expectation. We use the method for finding the expected value of a ratio, in Chapter 4, to get

$$\begin{aligned} E\left[\mathrm{RR}_{1w}\right] &= E\left[\frac{\sum_{i=1}^N Z_i R_i w_i x_{ic}}{\sum_{i=1}^N Z_i w_i x_{ic}} \right] \\ &= E\left[\frac{\sum_{i=1}^N Z_i R_i w_i x_{ic}}{\sum_{i=1}^N x_{ic}} \left(1 - \frac{\sum_{i=1}^N Z_i w_i x_{ic} - \sum_{i=1}^N x_{ic}}{\sum_{i=1}^N Z_i w_i x_{ic}} \right) \right] \\ &= E\left[\frac{\sum_{i=1}^N Z_i R_i w_i x_{ic}}{\sum_{i=1}^N x_{ic}} \right] - E\left[\frac{\left(\sum_{i=1}^N Z_i R_i w_i x_{ic}\right)\left(\sum_{i=1}^N Z_i w_i x_{ic} - \sum_{i=1}^N x_{ic}\right)}{\sum_{i=1}^N Z_i w_i x_{ic}} \right]. \end{aligned}$$

Note that if we define $P(R_i = 1 | Z_i = 1) = \phi_i$, we don't need to assume independence of Z_i and R_i (but this makes the notation in the chapter messier). In practice, of course a unit not selected for the sample doesn't get the opportunity to be a nonrespondent.

8.24 (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.

8.25 *R-indicators.* (a) When all response propensities are equal, the quantity under the square root equals 0, and $R(\boldsymbol{\phi}) = 1$, the largest value it can take on. Conversely, the only way to have $R(\boldsymbol{\phi}) = 1$ is for the variance of the ϕ_i to be 0.

The variance of the ϕ_i is largest when half are 0 and half are 1. Then

$$\sum_{i=1}^{N}\left(\phi_i - \frac{1}{N}\sum_{j=1}^{N}\phi_j\right)^2 = N/4$$

and $R(\boldsymbol{\phi}) = 1 - 2\sqrt{\frac{N}{4(N-1)}} \geq 0$.

(b) When $R(\boldsymbol{\phi}) = 1$, all response propensities are equal and the respondents are essentially a random subsample of the sample.

(c) We have

$$\begin{aligned}\frac{1}{N}\sum_{j\in\mathcal{S}} w_j\hat{\phi}_j &= \frac{1}{150104}\left[(30322)(0.6165) + (33013)(0.8525) + (27046)(0.9011) + (29272)(0.9613) + (30451)(1)\right] \\ &= 0.8647.\end{aligned}$$

Using the formula,

$$\begin{aligned}R(\hat{\boldsymbol{\phi}}) &= 1 - 2\sqrt{\frac{1}{N-1}\sum_{i\in\mathcal{S}} w_i\left(\hat{\phi}_i - \frac{1}{N}\sum_{j\in\mathcal{S}} w_j\hat{\phi}_j\right)^2} \\ &= 1 - 2\sqrt{\frac{1}{150{,}103}\left[30322(0.6165 - 0.8647)^2 + \ldots + 29272(0.9613 - 0.8647)^2 + 30451(1 - .8647)^2\right]} \\ &= 1 - 2\sqrt{0.0182495} \\ &= 0.73.\end{aligned}$$

(d) One example uses a model for ϕ where everyone is predicted to have the same response propensity, so that $R(\hat{\boldsymbol{\phi}}) = 1$, but the true ϕ_i's vary among units.

Chapter 9

Variance Estimation in Complex Surveys

9.1 The random group method would not be a good choice for this problem because only 5 districts were chosen from each region. This would result in a variance estimate with 4 degrees of freedom. Balanced repeated replication is possible, but modifications would need to be made to force the design into a structure with 2 psus per stratum. All of the other 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 (9.9),

$$\hat{V}_{\mathrm{JK}}(\bar{y}) = \frac{29}{30}\sum_{j=1}^{30}[\bar{y}_{(j)} - \bar{y}]^2 = 0.5326054.$$

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].

The SAS code below gives approximately the same interval:

```
proc surveymeans data=srs30 total=100 percentile=(25 50 75) nonsymcl;
   weight wt;
   var y;
run;
```

```
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
```

9.4 Here are estimates from SAS software. Other software packages may give slightly different estimates or CIs.

Percentile	Estimate	Std Error	95% Confidence Limits	
25 Q1	82980	8774	65712.560	100247.108
50 Median	194852	11415	172386.984	217317.195
75 Q3	373561	26935	320553.082	426569.638

9.5 The jackknife weights can be constructed in either R or SAS software.

9.6 Here is code in SAS software for creating the replicate stratified samples:

```
DATA public_college;
   set datalib.college;
   if control = 2 then delete; /* Remove the non-public schools from the data set */
   if ccsizset >= 15 then largeinst = 1;
   else if 6 <= ccsizset <= 14 then largeinst = 0;


proc sort data=public_college;
   by largeinst;

PROC SURVEYSELECT DATA = public_college method=srs n=10 reps=5 (repname = replicate) seed
    =38957273 stats out=collegereps;
   strata largeinst / alloc = prop;

/* Check for missing data */

proc means data=collegereps mean nmiss;
      class replicate;
          var tuitionfee_in tuitionfee_out;

proc print data=collegereps;
   var replicate largeinst instnm ugds tuitionfee_in tuitionfee_out samplingweight
    selectionprob;

proc freq data=collegereps;
   tables instnm;

proc sort data=collegereps;
   by replicate;

proc surveymeans data=collegereps PLOTS=none mean;
   by replicate;
   weight samplingweight;
   var tuitionfee_in tuitionfee_out;
   ratio 'out-of-state to in-state' tuitionfee_out / tuitionfee_in;
   ods output STATISTICS = Outstats RATIO = Ratiostats;
   title 'Calculate statistics from each replicate sample';
```

```
proc print data=Outstats;
proc print data=Ratiostats;

proc surveymeans data=Ratiostats mean var clm df;
   var Ratio;
   title 'Calculate the mean of the 5 ratios from the replicate samples, with CI';
run;
```

9.7

Variable	Random Group CI	Linearization CI
Avg age at first arrest	[12.85, 13.28]	[12.89, 13.20]
Median age, first arrest	[12.62, 12.91]	[12.57, 12.94]
Proportion violent offense	[0.388, 0.503]	[0.393, 0.493]
Proportion both parents	[0.275, 0.328]	[0.278, 0.326]
Proportion male	[0.890, 0.982]	[0.896, 0.966]
Proportion illegal drugs	[0.785, 0.872]	[0.794, 0.863]

The linearization CIs are slightly smaller because the variance estimator has more df.

9.8 The mean age is 16.64 with standard error 0.13 and 95% CI [16.38, 16.89].

9.9 From Exercise 4.3, $\hat{B} = 11.41946$, $\hat{\bar{y}}_r = t_x\hat{B} = 10.3\hat{B} = 117.6$, and $\text{SE}\,(\hat{\bar{y}}_r) = 3.98$. Using the jackknife, we have $\hat{B}_{(\cdot)} = 11.41937$, $\hat{\bar{y}}_{r(\cdot)} = 117.6$, and $\text{SE}\,(\hat{\bar{y}}_r) = 10.3\sqrt{0.1836} = 4.41$. The jackknife standard error is larger, partly because it does not include the fpc.

9.10 We use

$$\hat{V}_{\text{JK}}(\hat{\bar{y}}_r) = \frac{n-1}{n}\sum_{j=1}^{10}(\hat{\bar{y}}_{r(j)} - \hat{\bar{y}}_r)^2.$$

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,

$$\hat{V}_{JK}(\hat{\bar{y}}_r) = \frac{9}{10}\sum_{j=1}^{10}(\hat{\bar{y}}_{r(j)} - 0.4402907)^2 = 0.00526$$

9.11 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 $3.071 \times 10^{-5}/\sqrt{1 - \frac{300}{3078}} = 3.232 \times 10^{-5}$.

9.12 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	13	14	15	16
$\hat{F}(y)$	.3917	.4000	.4167	.4333	.5167	.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.

9.16 (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 the result is much more complicated for an SRS without replacement. Results from Finucan et al. (1974) may be used to find the moments.

9.17 (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,

$$\begin{aligned}
&\hat{B}_1 - B_1 \\
&\approx \frac{\partial h}{\partial a}(\hat{t}_{xy} - t_{xy}) + \frac{\partial h}{\partial b}(\hat{t}_x - t_x) + \frac{\partial h}{\partial c}(\hat{t}_y - t_y) \\
&\qquad + \frac{\partial h}{\partial d}(\hat{t}_{x^2} - t_{x^2}) + \frac{\partial h}{\partial e}(\hat{N} - N) \\
&= \frac{N}{Nt_{x^2} - (t_x)^2}\left[\hat{t}_{xy} - t_{xy} - (B_0 - B_1\bar{x}_U)(\hat{t}_x - t_x)\right.
\end{aligned}$$

$$
\begin{aligned}
& -\bar{x}_U(\hat{t}_y - t_y) - B_1(\hat{t}_{x^2} - t_{x^2}) + B_0\bar{x}_U(\hat{N} - N)\Big] \\
= & \frac{N}{Nt_{x^2} - (t_x)^2}\left[\sum_{i\in\mathcal{S}} w_i\left\{x_iy_i - (B_0 - B_1\bar{x}_U)x_i - \bar{x}_Uy_i - B_1x_i^2 + B_0\bar{x}_U\right\}\right] \\
& -\frac{N}{Nt_{x^2} - (t_x)^2}\left[t_{xy} - t_x(B_0 - B_1\bar{x}_U) - \bar{x}_Ut_y - B_1t_{x^2} + B_0N\bar{x}_U\right] \\
= & \frac{N}{Nt_{x^2} - (t_x)^2}\sum_{i\in\mathcal{S}} w_i\left(y_i - B_0 - B_1x_i\right)\left(x_i - \bar{x}_U\right).
\end{aligned}
$$

9.18 (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

$$
\begin{aligned}
q_i &= \frac{\partial h}{\partial t_1}y_i^2 + \frac{\partial h}{\partial t_2}y_i + \frac{\partial h}{\partial t_3} \\
&= \frac{1}{t_3 - 1}y_i^2 - 2\frac{t_2}{t_3(t_3 - 1)}y_i - \frac{1}{(t_3 - 1)^2}\left(t_1 - \frac{t_2}{t_3}\right) + \frac{1}{t_3 - 1}\frac{t_2}{t_3^2} \\
&= \frac{1}{t_3 - 1}\left(y_i^2 - 2\frac{t_2}{t_3}y_i - \frac{1}{(t_3 - 1)}\left(t_1 - \frac{t_2}{t_3}\right) + \frac{t_2}{t_3^2}\right)
\end{aligned}
$$

9.19 (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:

$$\begin{aligned}
\frac{\partial h}{\partial a} &= \frac{1}{\sqrt{(fc-a^2)(fe-b^2)}}\left(-b+\frac{a(fd-ab)}{fc-a^2}\right) \\
&= \frac{-t_y}{N(N-1)S_xS_y}+\frac{t_xR}{N(N-1)S_x^2} \\
\frac{\partial h}{\partial b} &= \frac{1}{\sqrt{(fc-a^2)(fe-b^2)}}\left(-a+\frac{b(fd-ab)}{fe-b^2}\right) \\
&= \frac{-t_y}{N(N-1)S_xS_y}+\frac{t_yR}{N(N-1)S_y^2} \\
\frac{\partial h}{\partial c} &= -\frac{1}{2\sqrt{(fc-a^2)(fe-b^2)}}\left(\frac{f(fd-ab)}{fc-a^2}\right) \\
&= -\frac{R}{2}\frac{1}{(N-1)S_x^2} \\
\frac{\partial h}{\partial d} &= \frac{f}{\sqrt{(fc-a^2)(fe-b^2)}} \\
\frac{\partial h}{\partial e} &= -\frac{1}{2\sqrt{(fc-a^2)(fe-b^2)}}\left(\frac{f(fd-ab)}{fe-b^2}\right) \\
&= -\frac{R}{2}\frac{1}{(N-1)S_y^2} \\
\frac{\partial h}{\partial f} &= \frac{d}{\sqrt{(fc-a^2)(fe-b^2)}}-\frac{fd-ab}{2\sqrt{(fc-a^2)(fe-b^2)}}\left(\frac{e}{fe-b^2}+\frac{c}{fc-a^2}\right) \\
&= \frac{t_{xy}}{N(N-1)S_xS_y}-\frac{R}{2}\left(\frac{t_{y^2}}{N(N-1)S_y^2}+\frac{t_{x^2}}{N(N-1)S_x^2}\right)
\end{aligned}$$

Then, by linearization,

$$\begin{aligned}
\hat{R}-R &\approx \frac{\partial h}{\partial a}(\hat{t}_x-t_x)+\frac{\partial h}{\partial b}(\hat{t}_y-t_y)+\frac{\partial h}{\partial c}(\hat{t}_{x^2}-t_{x^2}) \\
&\quad +\frac{\partial h}{\partial d}(\hat{t}_{xy}-t_{xy})+\frac{\partial h}{\partial e}(\hat{t}_{y^2}-t_{y^2})+\frac{\partial h}{\partial f}(\hat{N}-N) \\
&= \frac{1}{N(N-1)S_xS_y}\left[\left(-t_y+\frac{t_xRS_y}{S_x}\right)(\hat{t}_x-t_x)+(-t_x+\frac{t_yRS_x}{S_y})(\hat{t}_y-t_y)\right] \\
&\quad -\frac{R}{2}\frac{1}{(N-1)S_x^2}(\hat{t}_{x^2}-t_{x^2})+\frac{1}{(N-1)S_xS_y}(\hat{t}_{xy}-t_{xy}) \\
&\quad -\frac{R}{2}\frac{1}{(N-1)S_y^2}(\hat{t}_{y^2}-t_{y^2}) \\
&\quad +\left[\frac{t_{xy}}{N(N-1)S_xS_y}-\frac{R}{2}\left(\frac{t_{y^2}}{N(N-1)S_y^2}+\frac{t_{x^2}}{N(N-1)S_x^2}\right)\right](\hat{N}-N) \\
&= \frac{1}{N(N-1)S_xS_y}\left[\left(-t_y+\frac{t_xRS_y}{S_x}\right)(\hat{t}_x-t_x)+\left(-t_x+\frac{t_yRS_x}{S_y}\right)(\hat{t}_y-t_y)\right.
\end{aligned}$$

$$-\frac{NRS_y}{2S_x}(\hat{t}_{x^2}-t_{x^2})+N(\hat{t}_{xy}-t_{xy})-\frac{NRS_x}{2S_y}(\hat{t}_{y^2}-t_{y^2})$$
$$+\left\{t_{xy}-\frac{R}{2}\left(\frac{t_{y^2}S_x}{S_y}+\frac{t_{x^2}S_y}{S_x}\right)\right\}(\hat{N}-N)\Bigg]$$

This is somewhat easier to do in matrix terms. Let

$$\boldsymbol{\delta}=\left[\left(-\bar{y}_U+\frac{\bar{x}_U RS_y}{S_x}\right),\left(-\bar{x}_U+\frac{\bar{y}_U RS_x}{S_y}\right),-\frac{RS_y}{2S_x},-\frac{RS_x}{2S_y},1,\frac{t_{xy}}{N}-\frac{R}{2N}\left(\frac{t_{y^2}S_x}{S_y}+\frac{t_{x^2}S_y}{S_x}\right)\right]^T,$$

then

$$\mathrm{Cov}\,(\hat{R})\approx\frac{1}{[(N-1)S_xS_y]^2}\boldsymbol{\delta}^T\mathrm{Cov}\,(\hat{\mathbf{t}})\boldsymbol{\delta}.$$

9.20 (a) In the notation of Example 9.2, $\hat{B}=\left(\sum_{i\in\mathcal{S}}\frac{y_i}{\pi_i}\right)\Big/\left(\sum_{i\in\mathcal{S}}\frac{x_i}{\pi_i}\right)$. Define $d_i=y_i-Bx_i$, where $B=t_y/t_x=\bar{y}_{\mathcal{U}}$. From (9.3) in SDA,

$$E[(\hat{B}-B)^2]\approx\frac{1}{t_x^2}V[\hat{t}_{dHT}].$$

Exercise 27 of Chapter 6 gives the variance of the Horvitz-Thompson estimator for Poisson sampling, so that

$$\begin{aligned}V(\hat{t}_{yr})&=t_x^2E[(\hat{B}-B)^2]\\&\approx V[\hat{t}_{dHT}]\\&=V\left[\sum_{i=1}^N\frac{Z_i}{\pi_i}d_i\right]\\&=\sum_{i=1}^N\frac{d_i^2}{\pi_i}(1-\pi_i)\\&=\sum_{i=1}^N\frac{(y_i-x_it_y/t_x)^2}{\pi_i}(1-\pi_i).\end{aligned}$$

The result follows because $\hat{t}_{yr1}=t_x\hat{B}$ and hence $V(\hat{t}_{yr1})=t_x^2V(\hat{B})$.

(b) Note that

$$\hat{V}(\hat{t}_d)=\sum_{i=1}^N Z_i\frac{(1-\pi_i)d_i^2}{\pi_i^2}$$

so

The result follows because $\hat{B}$ is consistent for B and $\hat{t}_x$ is consistent for t_x.

(c) Letting $x_i = 1$ for all population units,

$$\hat{t}_{yr1} = \frac{N}{\hat{N}} \sum_{i \in \mathcal{S}} y_i/\pi_i$$

where $\hat{N} = \hat{t}_{x\text{HT}} = \sum_{i \in \mathcal{S}} 1/\pi_i$. Using part (a), $t_y/t_x = \bar{y}_{\mathcal{U}}$, so

$$V(\hat{t}_{yr1}) = \sum_{i=1}^{N} \frac{(y_i - \bar{y}_{\mathcal{U}})^2}{\pi_i} (1 - \pi_i) .$$

(d) Letting $x_i = \pi_i$ for all population units, we have $\hat{t}_x = m$, where $m = \sum_{i \in \mathcal{S}} \pi_i/\pi_i$ is the sample size for the Poisson sample. The population total is $t_x = E[m] = \sum_{i=1}^{N} E[Z_i] = \sum_{i=1}^{N} \pi_i$. We have

$$\hat{t}_{yr2} = \frac{E[m]}{m} \sum_{i \in \mathcal{S}} y_i/\pi_i$$

Using part (a), $t_y/t_x = t_y/E[m]$, so

$$\begin{aligned} V(\hat{t}_{yr2}) &= \sum_{i=1}^{N} \frac{(y_i - \pi_i t_y/E[m])^2}{\pi_i} (1 - \pi_i) \\ &= \sum_{i=1}^{N} \frac{1}{E[m]^2} \left(\frac{E[m] y_i}{\pi_i} - t_y \right)^2 \pi_i (1 - \pi_i) \end{aligned}$$

(e) We have three estimators, and the variance of each has the form

$$\sum_{i=1}^{N} \frac{(e_i)^2}{\pi_i} (1 - \pi_i) .$$

For $\hat{t}_{y\text{HT}}$, $e_{iHT} = y_i$. For $\hat{t}_{yr1}$, $e_{i1} = y_i - \bar{y}_{\mathcal{U}}$. And for $\hat{t}_{yr2}$, $e_{i2} = y_i - \pi_i t_y/E[m]$.

(i) If $\pi_i = n/N$, then for estimator $\hat{t}_{yr2}$, $E[m] = \sum_{i=1}^{N} \pi_i = n$. We have:

$$V(\hat{t}_{yHT}) = \frac{N}{n} (1 - n/N) \sum_{i=1}^{N} y_i^2,$$

$$V(\hat{t}_{yr1}) = \frac{N}{n} (1 - n/N) \sum_{i=1}^{N} (y_i - \bar{y}_{\mathcal{U}})^2,$$

$$V(\hat{t}_{yr2}) = \frac{N}{n} (1 - n/N) \sum_{i=1}^{N} (y_i - \bar{y}_{\mathcal{U}})^2 .$$

For this situation, the second and third ratio estimators are the same. Both, however, are expected to have smaller variance than the HT variance because they are looking at the sum of squares of

differences from the mean, whereas the HT estimator variance contains the uncorrected sum of squares. We know that

$$\sum_{i=1}^{N}(y_i - \bar{y}_{\mathcal{U}})^2 = \sum_{i=1}^{N} y_i^2 - N\bar{y}_{\mathcal{U}}^2 < \sum_{i=1}^{N} y_i^2$$

because all $y_i > 0$.

(ii) If $y_i = k\pi_i$, then $t_y = k\sum_{i=1}^{N}\pi_i = kE[m]$. The residuals, then, are

$$e_{i2} = y_i - \pi_i t_y / E[m] = y_i - k\pi_i = 0.$$

Thus, if y_i is proportional to the selection probability, the ratio estimator in (d) has variance 0. You can't get smaller than that!

9.21 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

$$V(\hat{t}_{\text{post}}) \approx V\Big[\sum_{l=1}^{L}\Big(\hat{t}_l - \frac{t_l}{N_l}\hat{N}_l\Big)\Big].$$

9.22 From (9.6),

$$\hat{V}_2(\hat{\theta}) = \frac{1}{R}\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \hat{\theta})^2.$$

Without loss of generality, let $\bar{y}_{\mathcal{U}} = 0$. 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

$$\begin{aligned} E\left[\frac{1}{R(R-1)}\sum_{r=1}^{R}(\bar{y}_r-\bar{y})^2\right] &= \frac{1}{R(R-1)}\sum_{r=1}^{R}E\left[\bar{y}_r^2-(\bar{y})^2\right] \\ &= \frac{1}{R(R-1)}\sum_{r=1}^{R}\left[V(\bar{y}_r)-V(\bar{y})\right] \\ &= \frac{1}{R(R-1)}\sum_{r=1}^{R}\left[\frac{S^2}{m}-\frac{S^2}{n}\right] \\ &= \frac{1}{R(R-1)}\sum_{r=1}^{R}\left[R\frac{S^2}{n}-\frac{S^2}{n}\right] \\ &= \frac{S^2}{n}. \end{aligned}$$

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 et al. (1953, pp. 99-100). They note that

$$\begin{aligned}\bar{y}_r^4 &= \frac{1}{m^4}\left[\sum_{i\in\mathcal{S}_r}y_i^4+4\sum_{i\neq j}y_i^3y_j+3\sum_{i\neq j}y_i^2y_j^2\right. \\ &\qquad \left.+6\sum_{i\neq j\neq k}y_i^2y_jy_k+\sum_{i\neq j\neq k\neq l}y_iy_jy_ky_l\right]\end{aligned}$$

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.

9.23 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 $\ell \neq h$.

9.24 *Fay's variation of BRR.* (a) When $\varepsilon = 0$,

$$w_i(\boldsymbol{\alpha}_r, \varepsilon) = \begin{cases} (1+\alpha_{rh})w_i & \text{if observation unit } i \text{ is in psu 1} \\ (1-\alpha_{rh})w_i & \text{if observation unit } i \text{ is in psu 2.} \end{cases},$$

which is the same weight as found in (9.7).

(b) The argument is similar to that in the previous exercise.

In stratified random sampling with $\hat{\theta} = \sum_{h=1}^{H}\frac{N_h}{N}\frac{y_{h1}+y_{h2}}{2}$,

$$\begin{aligned}
\hat{\theta}(\alpha_r, \varepsilon) &= \frac{\sum_{h=1}^{H}\sum_{i=1}^{2}w_{hi}(\alpha_r,\varepsilon)y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2}w_{hi}(\alpha_r,\varepsilon)} \\
&= \frac{\sum_{h=1}^{H}\frac{N_h}{2}[(1+\alpha_{rh}-\alpha_{rh}\varepsilon)y_{h1} + (1-\alpha_{rh}+\alpha_{rh}\varepsilon)y_{h2}]}{\sum_{h=1}^{H}\frac{N_h}{2}(2-\varepsilon+\varepsilon)} \\
&= \frac{\sum_{h=1}^{H}\frac{N_h}{2}[(1+\alpha_{rh}-\alpha_{rh}\varepsilon)y_{h1} + (1-\alpha_{rh}+\alpha_{rh}\varepsilon)y_{h2}]}{N}.
\end{aligned}$$

and

$$N\left[\hat{\theta}(\alpha_r,\varepsilon) - \hat{\theta}\right] = \sum_{h=1}^{H}\frac{N_h}{2}[(1+\alpha_{rh}-\alpha_{rh}\varepsilon)y_{h1} + (1-\alpha_{rh}+\alpha_{rh}\varepsilon)y_{h2}] - \sum_{h=1}^{H}\frac{N_h}{2}[y_{h1}+y_{h2}]$$

$$= \sum_{h=1}^{H} \frac{N_h}{2}[\alpha_{rh}(1-\varepsilon)y_{h1} - \alpha_{rh}(1-\varepsilon)y_{h2}]$$
$$= \sum_{h=1}^{H} \frac{N_h}{2}\alpha_{rh}(1-\varepsilon)[y_{h1} - y_{h2}]$$

$$\begin{aligned}
\hat{V}_{\text{BRR},\varepsilon}(\hat{\theta}) &= \frac{1}{N^2 R(1-\varepsilon)^2}\sum_{r=1}^{R}\left[\hat{\theta}(\alpha_r,\varepsilon) - \hat{\theta}\right]^2 \\
&= \frac{1}{N^2 R(1-\varepsilon)^2}\sum_{r=1}^{R}\left[\sum_{h=1}^{H}\frac{N_h}{2}\alpha_{rh}(1-\varepsilon)[y_{h1}-y_{h2}]\right]^2 \\
&= \frac{1}{N^2 R(1-\varepsilon)^2}\sum_{r=1}^{R}\left[\sum_{h=1}^{H}\frac{N_h}{2}\alpha_{rh}(1-\varepsilon)[y_{h1}-y_{h2}]\right]\left[\sum_{l=1}^{H}\frac{N_l}{2}\alpha_{rl}(1-\varepsilon)[y_{l1}-y_{l2}]\right] \\
&= \frac{1}{N^2 R}\sum_{r=1}^{R}\left[\sum_{h=1}^{H}\frac{N_h}{2}\alpha_{rh}[y_{h1}-y_{h2}]\right]\left[\sum_{l=1}^{H}\frac{N_l}{2}\alpha_{rl}[y_{l1}-y_{l2}]\right] \\
&= \frac{1}{N^2 R}\sum_{r=1}^{R}\sum_{h=1}^{H}\sum_{l=1}^{H}\frac{N_h}{2}\frac{N_l}{2}\alpha_{rh}[y_{h1}-y_{h2}]\alpha_{rl}[y_{l1}-y_{l2}] \\
&= \frac{1}{N^2 R}\sum_{h=1}^{H}\sum_{l=1}^{H}\frac{N_h}{2}\frac{N_l}{2}[y_{h1}-y_{h2}][y_{l1}-y_{l2}]\sum_{r=1}^{R}\alpha_{rh}\alpha_{rl} \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\sum_{l=1}^{H}\frac{N_h^2}{4}(y_{h1}-y_{h2})^2 = \hat{V}_{\text{str}}(\bar{y}_{\text{str}}).
\end{aligned}$$

The last step follows because the replicates are balanced, so that $\sum_{r=1}^{R}\alpha_{rh}\alpha_{rl} = 0$ for $h \neq l$ and $\sum_{r=1}^{R}\alpha_{rh}^2 = R$.

9.25 Answers will vary depending on the half-samples selected.

9.26 As noted in SDA,

$$\hat{V}_{\text{str}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{(y_{h1}-y_{h2})^2}{4}.$$

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,

$$\begin{aligned}&\frac{1}{2R}\sum_{r=1}^{R}\{[\hat{\theta}(\alpha_r) - \hat{\theta}]^2 + [\hat{\theta}(-\alpha_r) - \hat{\theta}]^2\} \\ &= \frac{1}{2R}\sum_{r=1}^{R}\left\{\left[\sum_{h=1}^{H}\frac{\alpha_{rh}}{2}\frac{N_h}{N}(y_{h1}-y_{h2})\right]^2 + \left[\sum_{h=1}^{H}\frac{-\alpha_{rh}}{2}\frac{N_h}{N}(y_{h1}-y_{h2})\right]^2\right\} \\ &= \frac{1}{2R}\sum_{r=1}^{R}\sum_{h=1}^{H}\frac{\alpha_{rh}^2}{2}\left(\frac{N_h}{N}\right)^2(y_{h1}-y_{h2})^2 \\ &= \frac{1}{4}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(y_{h1}-y_{h2})^2.\end{aligned}$$

9.27 Note that

$$\hat{t}(\alpha_r) = \sum_{h=1}^{H} N_h\left[\frac{\alpha_{rh}}{2}(y_{h1}-y_{h2}) + \bar{y}_h\right] = \sum_{h=1}^{H}\frac{N_h a_{rh}}{2}(y_{h1}-y_{h2}) + \hat{t}$$

and

$$\begin{aligned}[\hat{t}(\alpha_r)]^2 = \sum_{h=1}^{H}\sum_{k=1}^{H}\frac{N_h N_k \alpha_{rh}\alpha_{rk}}{4}(y_{h1}-y_{h2})(y_{k1}-y_{k2}) \\ + 2\hat{t}\sum_{h=1}^{H}\frac{N_h\alpha_{rh}}{2}(y_{h1}-y_{h2}) + \hat{t}^2.\end{aligned}$$

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 $\frac{1}{4R} \sum_{r=1}^{R} [\hat{\theta}(\alpha_r) - \hat{\theta}(-\alpha_r)]^2$.

9.29 *Demnati-Rao approach.* 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,

$$\begin{aligned}
z_i &= \frac{\partial g(\mathbf{w}, \mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_L)}{\partial w_i} \\
&= \sum_{l=1}^{L} \left\{ \frac{N_l x_{li} y_i}{\sum_{j \in \mathcal{S}} w_j x_{lj}} - \frac{N_l x_{li} \sum_{j \in \mathcal{S}} w_j x_{lj} y_j}{\left(\sum_{j \in \mathcal{S}} w_j x_{lj} \right)^2} \right\} \\
&= \sum_{l=1}^{L} \left\{ \frac{N_l x_{li} y_i}{\hat{N}_l} - \frac{N_l x_{li} \hat{t}_{yl}}{\hat{N}_l^2} \right\}
\end{aligned}$$

$$= \sum_{l=1}^{L} x_{li} \frac{N_l}{\hat{N}_l} \left(y_i - \frac{\hat{t}_{yl}}{\hat{N}_l} \right).$$

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.

9.30 Use the grouped jackknife code in Chapter 10 of Lohr (2022).

9.31 From Chapter 5,

$$\begin{aligned} V(\hat{t}) &\approx N^2 \frac{M\text{MSB}}{n} \\ &= \frac{N^2 M}{n} \frac{NM-1}{M(N-1)} S^2[1+(M-1)ICC] \\ &\approx \frac{NM}{n} \frac{NM}{M} p(1-p)[1+(M-1)ICC] \end{aligned}$$

Consequently, the relative variance v can be written as $\beta_0 + \beta_1/t$, where

$$\beta_0 = -\frac{1}{nM}[1+(M-1)ICC] + \frac{1}{nt}N[1+(M-1)ICC].$$

9.32 (a) From (9.3),

$$\begin{aligned} V[\hat{B}] &\approx E\left[\left\{ -\frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y) \right\}^2 \right] \\ &= \frac{t_y^2}{t_x^2} E\left[\left\{ -\frac{1}{t_x}(\hat{t}_x - t_x) + \frac{1}{t_y}(\hat{t}_y - t_y) \right\}^2 \right] \\ &= \frac{t_y^2}{t_x^2} \left[\frac{V(\hat{t}_x)}{t_x^2} + \frac{V(\hat{t}_y}{t_y^2} - 2\text{Cov}\left(\frac{\hat{t}_x}{t_x}, \frac{\hat{t}_y}{t_y} \right) \right] \\ &= \frac{t_y^2}{t_x^2} \left[\frac{V(\hat{t}_x)}{t_x^2} + \frac{V(\hat{t}_y}{t_y^2} - 2\frac{B}{t_x t_y} V\left(\hat{t}_x \right) \right] \\ &= \frac{t_y^2}{t_x^2} \left[\frac{V(\hat{t}_y}{t_y^2} - \frac{V(\hat{t}_x)}{t_x^2} \right] \end{aligned}$$

Using the fitted model from (9.15),

$$\frac{\hat{V}(\hat{t}_x)}{\hat{t}_x^2} = a + b/\hat{t}_x$$

and

$$\frac{\hat{V}(\hat{t}_y)}{\hat{t}_y^2} = a + b/\hat{t}_y$$

Consequently, substituting estimators for the population quantities,

$$\hat{V}[\hat{B}] \;=\; \hat{B}^2 \left[a + \frac{b}{\hat{t}_y} - a - \frac{b}{\hat{t}_x} \right],$$

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 contain clustered or pseudoreplicated data. See Alf and Lohr (2007) for a review of how books ignore clustering in the data.

10.2 Answers will vary.

10.3 (a) Observed and expected (in parentheses) proportions are given in the following table:

		Abuse	
		No	Yes
Symptom	No	.7542	.1017
		(.7109)	(.1451)
	Yes	.0763	.0678
		(.1196)	(.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 χ^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 estimated population counts. A second-order Rao–Scott test has $F = 66.7$ and p-value (comparing to an F distribution with 1 and 9 df) < 0.0001. Other variations have similar p-values. These variables are strongly associated.

	mathlevel		
Estimated Total	1	2	
readlevel=1	8705	307	9012
readlevel=2	3599	4622	8261
	12303	4969	17273

10.6 (a) The following table gives the estimated population size in each cell, calculated using the weights.

	At Least 1 Female Detective?		
Estimated Total	Yes	No	
Male Author	49.25	398.75	448
Female Author	122.00	85.00	207
	171.25	483.75	655

(b) The first-order Rao-Scott test statistics are $X^2 = 15.43$ and $F = 11.52$. Comparing this to an F distribution with 1 and 48 df gives p-value = 0.0014. There is strong evidence that female authors are more likely to write novels with at least one female detective than are male authors. Note that other chi-square tests might be used. (These values are from SAS software; R gives slightly different values, but with the same conclusions, with F=11.84 and p-value = .002.)

10.7 (a) The following table gives the estimated population size in each cell, calculated using the weights

Estimated Total	Genre Private Eye	Procedural	Suspense	
Male Author	91.17	73.83	283.00	448
Female Author	55.50	30.17	121.33	207
	146.67	104.00	404.33	655

(b) The Rao-Scott test statistic is $X^2 = 0.2645$. Comparing this to an χ^2 distribution with 2 df gives p-value = 0.8761. There is no evidence of an association between author gender and genre.

10.8 (a) Here is the table:

Estimated Total	Historical? Yes	No	
Male Author	120.9	327.1	448
Female Author	47.7	159.3	207
	168.6	486.4	655

(b) $F = 0.09$ with 1 and 48 df, and p-value = 0.78, No evidence of association.

10.9 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 χ^2_1 distribution, obtaining a p-value of 0.4, which is larger than the p-value in Example 10.5.

10.10 (a) The estimated population contingency table is

Est Pop Count		ConfInt No	ConfInt Yes	
RandomSel	No	112.95	273.91	386.85
	Yes	41.26	118.88	160.15
		154.21	392.79	547.00

The Rao-Scott chi-square test statistic is 0.2370 with p-value = 0.63. There is no evidence of an association.

(b) Here is the estimated population contingency table.

Est Pop Count		authors_cat 0	authors_cat 1	
RandomSel	No	201.86	184.99	386.85
	Yes	121.64	38.51	160.15
		323.50	223.50	547.00

The odds ratio is 0.3454, with 95% [0.1723 0.6927]. Papers with more authors are less likely to have used random selection.

10.11 (a) The contingency table (for complete data) is as follows:

	Break again? No	Break again? Yes	
Faculty	65	167	232
Classified staff	55	459	514
Administrative staff	11	75	86
Academic professional	9	58	67
	140	759	899

$X_p^2 = 37.3$; comparing to a χ_3^2 distribution gives ρ-value $< .0001$. We can use the χ^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, if you used the reverse categories, 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	4.04×10^{-5}
C.S.	1960	514	−.059	0.0111	4.52×10^{-6}
A.S.	252	86	−.061	0.0207	7.42×10^{-7}
A.P.	95	66	−.076	0.0349	1.08×10^{-7}
Total	3681	894			4.58×10^{-5}

Thus $\hat{V}_L(\hat{\theta}) = 4.58 \times 10^{-5}$ and

$$X_W^2 = \frac{\hat{\theta}^2}{\hat{V}_L(\hat{\theta})} = \frac{0.00165}{4.58 \times 10^{-5}} = 36.2.$$

We reject the null hypothesis with p-value < 0.0001.

10.12 Answers will vary, depending on how the categories for zprep are formed.

10.13 (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 χ_4^2 distribution gives p-value= .001.

(b) Using (10.9), $E[X^2] \approx E[G^2] \approx 6.84$

(c) $X_F^2 = G_F^2 = \dfrac{4X^2}{6.84} = 10.7$, with p-value = .03 (comparing to a χ_4^2 distribution).

10.14 (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 χ^2 distribution with $(2-1)(5-1) = 4$ *df* gives "*p*-values" that are less than 1×10^{-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 1.0×10^{-9}.

(e) The *p*-value for X_s^2 is 2.6×10^{-8}, still very small.

10.15 Here are calculations from R:

```
data(syc)
syc$ageclass <- syc$age
syc$ageclass[syc$age <= 15] <- 1
syc$ageclass[syc$age == 16 | syc$age == 17] <- 2
syc$ageclass[18 <= syc$age] <- 3
syc$currprop <- syc$crimtype - 1
syc$currprop[syc$crimtype != 2 | is.na(syc$crimtype)] <- 0
table(syc$currprop,syc$crimtype)  # check variable construction
```

```
dsyc<-svydesign(ids=~psu,weights=~finalwt,strata=~stratum,nest=TRUE,data=syc)
svytable(~currprop+ageclass,design=dsyc)/sum(syc$finalwt)
##          ageclass
## currprop          1          2          3
##        0 0.14505038 0.24360307 0.18842955
##        1 0.13549496 0.20306253 0.08435951

svychisq(~currprop+ageclass, design=dsyc)

##         Pearson's X^2: Rao & Scott adjustment

## data:  svychisq(~currprop + ageclass, design = dsyc)
## F = 11.461, ndf = 1.7148, ddf = 1449.0457, p-value = 3.723e-05
```

10.16 Here are the estimated probabilities and design effects:

```
Estimated Probability
        ageclass
bmiclass          1          2          3     Total
       1 0.12576233 0.09067267 0.06676095  0.283196
       2 0.10584261 0.11549602 0.09740231  0.318741
       3 0.12998858 0.15383369 0.11424085  0.398063
   Total 0.3615935  0.3600024  0.2784041
```

```
Design Effect
        ageclass
bmiclass    1      2     3   Total
       1 2.07   5.75  3.16   6.08
       2 2.81   3.75  3.35   0.87
       3 3.22   1.47  3.60   5.42
   Total 4.17   3.24  5.84
```

The second-order Rao–Scott test statistic is $F = 7.98$, which is compared to an F distribution with 3.36 and 50.38 df, giving a p-value 0.0001.

10.17 Estimated probabilities are:

```
         ageclass
cholclass          1          2          3
        0 0.25991356 0.17613118 0.16531674
        1 0.09708877 0.18565141 0.11589835
```

The second-order Rao–Scott test statistic is $F = 105.7408$, which is compared to an F distribution with 1.97 and 29.6 df, giving a p-value < 0.0001.

10.19 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 χ^2_1 distribution because of the clustering.

10.20 (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,

$$\begin{aligned} V_L(\hat{\theta}) = p_{22}^2 \frac{p_{11}(1-p_{11})}{n} + \cdots + p_{21}^2 \frac{p_{12}(1-p_{12})}{n} \\ - 2\frac{p_{11}^2 p_{22}^2}{n} + 2\frac{p_{11}p_{22}p_{12}p_{21}}{n} + \cdots - 2\frac{p_{12}^2 p_{21}^2}{n} \end{aligned}$$

$$= \frac{1}{n}\left\{-4p_{11}^2p_{22}^2 - 4p_{12}^2p_{21}^2 + 8p_{11}p_{22}p_{12}p_{21} + p_{11}p_{22}(p_{11}+p_{22}) + p_{12}p_{21}(p_{12}+p_{21})\right\}.$$

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,

$$X_W^2 = n\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}} = X_p^2.$$

10.21 (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

$$\begin{aligned}V_L(\hat{\theta}) &= V\left(\frac{\hat{p}_{11}}{p_{11}} + \frac{\hat{p}_{22}}{p_{22}} - \frac{\hat{p}_{12}}{p_{12}} - \frac{\hat{p}_{21}}{p_{21}}\right) \\ &= \sum_{j=1}^{2}\sum_{k=1}^{2}\frac{1}{p_{jk}^2}V(\hat{p}_{jk}) \\ &\quad + \frac{2}{p_{11}p_{22}}\,\mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{22}) - \frac{2}{p_{11}p_{12}}\,\mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{12}) \\ &\quad - \frac{2}{p_{11}p_{21}}\,\mathrm{Cov}\,(\hat{p}_{11}, \hat{p}_{21}) - \frac{2}{p_{12}p_{22}}\,\mathrm{Cov}\,(\hat{p}_{22}, \hat{p}_{12})\end{aligned}$$

$$-\frac{2}{p_{22}p_{21}}\,\mathrm{Cov}\,(\hat{p}_{22},\hat{p}_{21})+\frac{2}{p_{12}p_{21}}\,\mathrm{Cov}\,(\hat{p}_{12},\hat{p}_{21}). \tag{10.1}$$

(b) Under multinomial sampling,

$$V_L(\hat{\theta})=\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{p_{jk}(1-p_{jk})}{np_{jk}^2}+\frac{4}{n}=\frac{1}{n}\Big(\frac{1}{p_{11}}+\frac{1}{p_{12}}+\frac{1}{p_{21}}+\frac{1}{p_{22}}\Big)$$

and

$$\hat{V}_L(\hat{\theta})=\frac{1}{n}\Big(\frac{1}{\hat{p}_{11}}+\frac{1}{\hat{p}_{12}}+\frac{1}{\hat{p}_{21}}+\frac{1}{\hat{p}_{22}}\Big)=\frac{1}{x_{11}}+\frac{1}{x_{12}}+\frac{1}{x_{21}}+\frac{1}{x_{22}}.$$

This is the estimated variance given in Section 10.1.1.

10.22 The test statistic is $\hat{\eta}/\sqrt{\hat{V}_L(\hat{\eta})}$, where $\hat{V}_L(\hat{\eta})$ is given in Equation (SM10.1).

10.24 In a multinomial sample, all design effects are 1. From (10.9), under H_0

$$\begin{aligned}E[X^2]&=\sum_{i=1}^{r}\sum_{j=1}^{c}(1-p_{ij})-\sum_{i=1}^{r}(1-p_{i+})-\sum_{j=1}^{c}(1-p_{+j})\\&=rc-1-(r-1)-(c-1)\\&=(r-1)(c-1).\end{aligned}$$

10.25 *Rao-Scott tests.* (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

$$\mathbf{Y}^T\mathbf{C}\mathbf{Y}=\mathbf{U}^T\boldsymbol{\Lambda}\mathbf{U}=\sum_{i=1}^{k}\lambda_iU_i^2.$$

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 $W_i=U_i^2\sim\chi_1^2$ 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 $\sum\lambda_iW_i$, where the λ_i's are the eigenvalues of

$$\mathbf{V}(\hat{\boldsymbol{\theta}})^{1/2}\mathbf{A}^{-1}\mathbf{V}(\hat{\boldsymbol{\theta}})^{1/2}.$$

(c)

$$E[\hat{\boldsymbol{\theta}}^T \mathbf{A}^{-1} \hat{\boldsymbol{\theta}}] \approx \sum \lambda_i,$$

$$V[\hat{\boldsymbol{\theta}}^T \mathbf{A}^{-1} \hat{\boldsymbol{\theta}}] \approx 2 \sum \lambda_i,$$

10.26 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.

10.27 Both test statistics are very large. We obtain the first-order Rao-Scott χ^2 statistic is 2851, while the Wald test statistic is 4662. There is strong evidence that the variables are associated.

Chapter 11

Regression with Complex Survey Data

11.1 Answers will vary.

11.2 Answers will vary.

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 the SURVEYREG procedure, we get $\bar{y}_1 - \bar{y}_2 = 15.27$ with 95% CI [7.625, 22.918]. Since 0 is not in the CI, there is evidence that the domain means differ.

11.4 (a) Answers will vary, since students will generate their own samples. The lines should be close to the fitted regression line for the truncated population, which is

$$\hat{y} = 43.79 + 1.789x.$$

The line is much flatter than the one in Figure 11.4.

The following code can be used with SAS software to draw the sample.

```
proc sort data=anthrop;
   by  height;

data anthrop1; /* Keep the lowest 2000 values in the data set */
   set anthrop;
   if _N_ <= 2000;

proc reg data = anthrop1;
   model height = finger;
   output out=regpred pred=predicted residual=resid;
```

```
   title 'Truncated population regression line';

proc surveyselect data=anthrop1 method=srs n=200 out=srs_lowheight seed = 2384725 stats;
run;
```

(b) We use exactly the same code as before, except now we sort the data by finger instead of by height. The population line is

$$\hat{y} = 31.20 + 2.96x.$$

These values of the slope and intercept are quite close to the values given in Figure 11.4. Student samples should obtain slopes and intercepts somewhat close to these. The standard errors will be larger than from an SRS of size 200 from the full population, however, reflecting the lesser number of data points and the reduced spread of the x's.

(c) Why the difference? Regression predicts y from x. The second population reduces the range of x, but the data set has the same basic relationship across the range so the regression equation is about the same. The first population, however, eliminates the large values of y, regardless of their x values. This distorts the relationship for the points that remain.

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 (a) The means for the two domains are as follows:

gender	N	Mean	Std Error	95% CL for Mean	
F	32	21.356250	1.764618	17.7058603	25.0066397
M	74	22.214065	1.117597	19.9820697	24.4460608

(b) The men minus women difference is 0.857815, with 95% CI [-3.46, 5.17]. There is no significant difference in timelag for the genders. Note, however, the outlier value of 50 years for one man in the sample.

(c) They are independent because *gender* is a stratification variable and sampling was done independently in different strata.

11.6 We obtain $\hat{B}_0 = 14.2725$ and $\hat{B}_1 = 0.08138$. Using Equation (11.9), 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

$$\begin{aligned}\hat{V}_L(\hat{B}_1) &= \hat{V}\left(\sum w_i q_i\right) \Big/ \left[\sum w_i x_i^2 - \frac{(\sum w_i x_i)^2}{\sum w_i}\right]^2 \\ &= \hat{V}\left(\frac{\sum w_i q_i}{\sum w_i x_i^2 - (\sum w_i x_i)^2 / \sum w_i}\right) = 0.000261\end{aligned}$$

$$SE_L(\hat{B}_1) = .016.$$

Here is output from SAS software:

```
data nybight;
   set datadir.nybight;
   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

                            Standard
Parameter      Estimate         Error    t Value    Pr > |t|

Intercept    14.2725002    2.13426397       6.69      <.0001
catchnum      0.0813836    0.01612252       5.05      <.0001
```

11.8 Using the weights as in Exercise 11.6, the estimated regression coefficients are $\hat{B}_0 = 7.569$ and $\hat{B}_1 = 0.0778$. From equation (11.9), $\hat{V}_L(\hat{B}_1) = 0.068$, (alternatively, $\hat{V}_{JK}(\hat{B}_1) = 0.070$). The slope is not significantly different from 0. Here is output from SAS software:

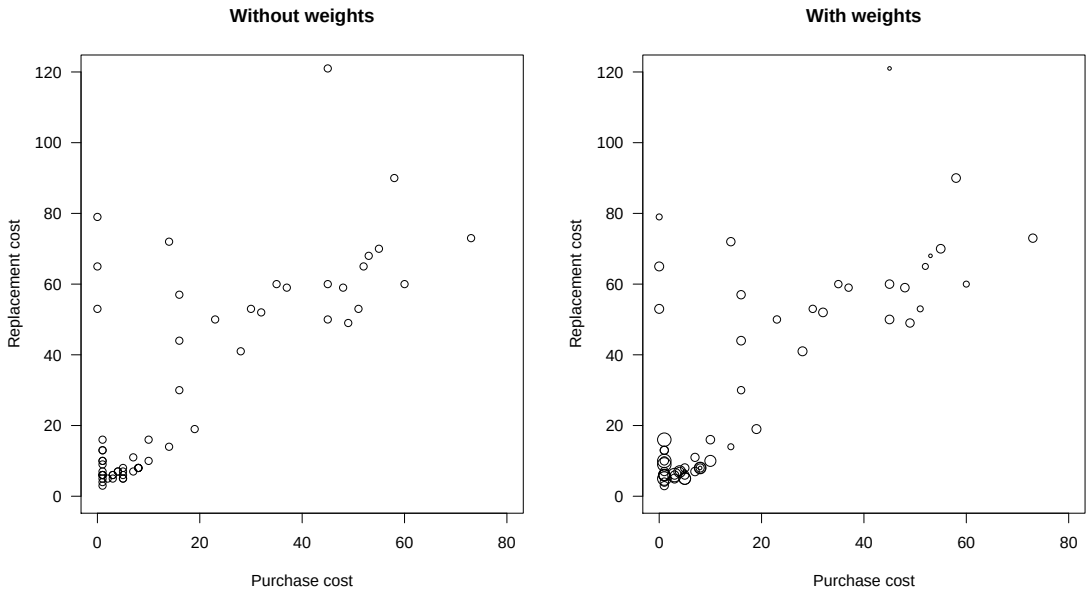
```
                        Standard                              95% Confidence
Parameter    Estimate       Error  t Value  Pr > |t|          Interval

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
```

11.9 (a) The design-based regression equation is *preprmin* $= 187.5 - 0.36$ (*size*), with standard errors 22 and 0.81. The slope is not significantly different from zero, and this analysis gives no indication that the variables are associated. Note, however, that the regression is driven by an outlier with respect to *size*.

(b) A model-based analysis would use a random-effects model with schools as the clusters. One would also want to include the district size as a fixed effect.

11.10 (a, b)



(c) Here is output from SAS software (without fpc because this is a regression analysis). SEs from R are slightly different (3.54 for intercept and 0.120 for slope).

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.

11.12 The regression parameters for predicting the probability that random selection was used are:

Parameter	Estimate	Lower CL	Upper CL
Intercept	0.0828	-0.7196	0.8852
NumAuthors	-0.1892	-0.3452	-0.0332

11.13 Here is code in SAS software.

```
proc surveyreg data=mysteries;
   class genre;
   weight p1weight;
   strata stratum;
   model fdetect = genre /solution clparm;
   lsmeans genre / diff adjust=tukey;
run;
```

The F statistic is 4.22, with p-value 0.02. This is evidence that the number of female detectives differ by genre. The LSMEANS statement tells which genres have the most and least female detectives: procedurals have the least, and are significantly different from suspense novels, with the most.

11.14 Answers will vary

11.15

Parameter	Estimate	Std Error	t Value	p-value	95% Confidence Interval	
Intercept	5.20949965	0.24427426	21.33	<.0001	4.68884139	5.73015790
bmxbmi	0.60099037	0.00786422	76.42	<.0001	0.58422819	0.61775254

11.16

Parameter	Estimate	Std Error	t Value	p-value	95% Confidence Interval	
Intercept	11.4587056	1.28439826	8.92	<.0001	8.72107557	14.1963357
bmxarmc	2.6697325	0.03754660	71.10	<.0001	2.58970377	2.7497611

11.17 Here is code in SAS software:

```
proc surveyreg data=nhanes nomcar;
  class raceeth riagendr;
  weight wtmec2yr;
  strata sdmvstra;
  cluster sdmvpsu;
  domain adult;
  model bmxbmi = raceeth riagendr raceeth*riagendr/ solution clparm;
  lsmeans raceeth*riagendr /diff adjust=tukey;
  title 'Compare domains using SURVEYREG procedure and class variables raceeth, riagendr'
    ;
run;
```

```
Tests of Model Effects
Effect      Num DF      F Value     Pr > F
Model            9       136.80     <.0001
```

```
Intercept          1      29695.8      <.0001
raceeth            4       136.25      <.0001
riagendr           1         7.30      0.0164
raceeth*riagendr   4        11.89      0.0001
```

Estimated means for groups and their differences are given below. Gender = 1 corresponds to male; gender = 2 corresponds to female. Note that these estimates differ slightly from those in Fryar et al. (2018) because they define groups using age at time of medical examination, which can include some extra people who were 19 at screening and turned 20 before the medical examination.

raceeth*riagendr Least Squares Means

raceeth	Gender	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Asian	1	25.3165	0.1662	15	152.29	<.0001	0.05	24.9622	25.6708
Asian	2	24.6676	0.2478	15	99.53	<.0001	0.05	24.1394	25.1959
Black	1	29.0344	0.3799	15	76.42	<.0001	0.05	28.2246	29.8442
Black	2	31.8280	0.4683	15	67.96	<.0001	0.05	30.8298	32.8262
Hispanic	1	30.0513	0.3674	15	81.78	<.0001	0.05	29.2681	30.8345
Hispanic	2	31.1296	0.3213	15	96.90	<.0001	0.05	30.4448	31.8143
Other	1	30.0696	0.8452	15	35.58	<.0001	0.05	28.2681	31.8712
Other	2	30.7896	0.5773	15	53.33	<.0001	0.05	29.5591	32.0201
White	1	29.1752	0.2709	15	107.70	<.0001	0.05	28.5979	29.7526
White	2	29.2822	0.3325	15	88.07	<.0001	0.05	28.5735	29.9909

11.18 Here is the code and output for the SURVEYLOGISTIC procedure in SAS software (the other statements are the same as for Example 11.12:

```
proc surveylogistic data=nhanes nomcar;
  weight wtmec2yr;
  strata sdmvstra;
  cluster sdmvpsu;
  domain adult;
  model bmi30 (event = '1') = bmxwaist female bmxwaist*female/ clparm clodds df = design;

  title 'Logistic regression treating female as numeric variable, with interaction';
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-29.8994	1.7362	-17.22	<.0001	-33.6000	-26.1988
bmxwaist	0.2804	0.0167	16.82	<.0001	0.2448	0.3159
female	1.4733	2.3783	0.62	0.5449	-3.5960	6.5425
bmxwaist*female	0.00102	0.0226	0.05	0.9645	-0.0471	0.0492

NOTE: The degrees of freedom for the t tests is 15.

11.19

Parameter	Estimate	Std Error	t Value	p-value	95% Confidence Interval	
Intercept	11.9791787	1.92112280	6.24	0.0002	7.63329700	16.3250604
math	0.5622974	0.04243710	13.25	<.0001	0.46629797	0.6582967

11.20 The estimated female minus male difference is 0.402 with SE 1.14 and 95% CI [-2.16, 2.97]. There is no evidence that the mean math score differs by gender.

11.21 From (11.5),

$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)\Big/N}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 /N} \\ &= \frac{N_1\bar{y}_{1U} - N_1\bar{y}_U}{N_1 - N_1^2/N} \\ &= \frac{\bar{y}_{1U} - (N_1\bar{y}_{1U} + N_2\bar{y}_{2U})/N}{1 - N_1/N} \\ &= \bar{y}_{1U} - \bar{y}_{2U}. \end{aligned}$$

From (11.6),

$$\begin{aligned} B_0 &= \frac{t_y - B_1 t_x}{N} \\ &= \bar{y}_U - B_1\bar{x}_U \\ &= \frac{N_1\bar{y}_{1U} + N_2\bar{y}_{2U}}{N} - \frac{N_1}{N}(\bar{y}_{1U} - \bar{y}_{2U}) \\ &= \bar{y}_{2U}. \end{aligned}$$

11.22 (a) Note that every member of the sample has data for the response indicator R_i. Using the survey-weighted regression parameters, we have

$$\begin{aligned} \hat{\mathbf{B}} &= \left[\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T\right]^{-1} \sum_{i\in\mathcal{S}} w_i \mathbf{x}_i R_i \\ &= \left[diag(\sum_{i\in\mathcal{S}} x_{i1} \ldots \sum_{i\in\mathcal{S}} x_{iC}\right]^{-1} \left[\sum_{i\in\mathcal{S}} x_{i1}R_i \ldots \sum_{i\in\mathcal{S}} x_{iC}R_i\right]^T \\ &= \left[\hat{\phi}_1 \ldots \hat{\phi}_C\right]^T \end{aligned}$$

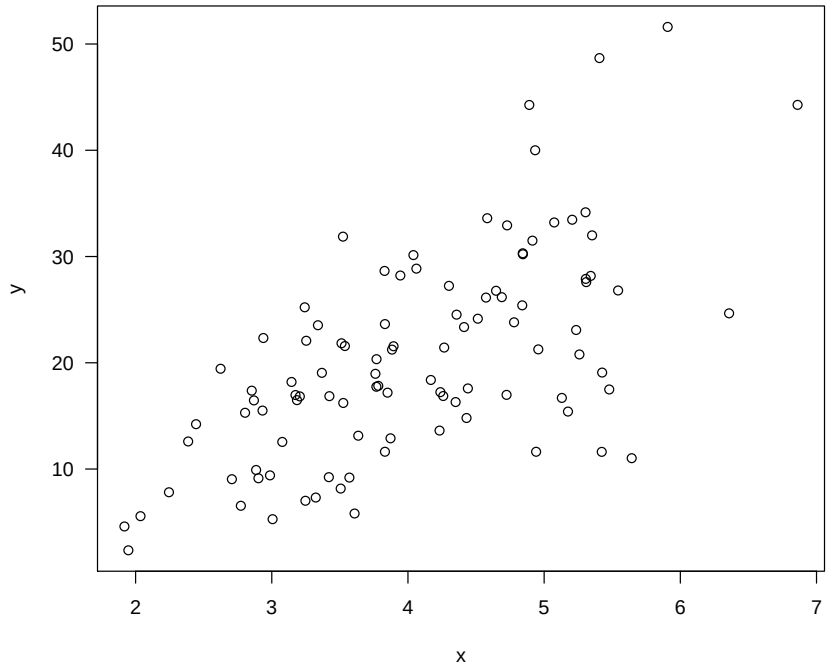
(b) Since R_i is 0 or 1, the numerator of $\hat{B}_c$ is always less than or equal to the denominator.

(c)

$$V(\hat{\mathbf{B}}) = \left(\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \hat{V} \left(\sum_{i\in\mathcal{S}} w_i \mathbf{q}_i\right) \left(\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1}$$
$$= \left[diag(\sum_{i\in\mathcal{S}} x_{i1} \ldots \sum_{i\in\mathcal{S}} x_{iC}\right]^{-1} \hat{V} \left(\sum_{i\in\mathcal{S}} w_i \mathbf{q}_i\right) \left[diag(\sum_{i\in\mathcal{S}} x_{i1} \ldots \sum_{i\in\mathcal{S}} x_{iC}\right]^{-1}$$

11.23 Answers will vary.

11.24 (a)



(b) $\hat{\beta}_0 = -4.096$; $\hat{\beta}_1 = 6.049$

(c) We estimate the model-based variance using the regression software: $\hat{V}_M(\hat{\beta}_1) = 0.541$.

$$\hat{V}_L(\hat{\beta}_1) = \frac{n\sum(x_i - \bar{x})^2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-1)[\sum(x_i - \bar{x})^2]^2} = 0.685$$

$\hat{V}_L$ is larger, as we would expect since the plot exhibits unequal variances.

11.25 Bayes' theorem implies that

$$\begin{aligned} f_{Y|Z}(y|z) &= \frac{f_Y(y) f_{Z|Y}(z|y)}{\int_{-\infty}^{\infty} f_Y(t) f_{Z|Y}(z|t) dt} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp[-(y - \mathbf{x}^T\boldsymbol{\beta})^2/(2\sigma^2)] \exp\left[\gamma_1 y + \gamma_2 y^2 + g(\mathbf{x}_i)\right]}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-(t - \mathbf{x}^T\boldsymbol{\beta})^2/(2\sigma^2)] \exp\left[\gamma_1 t + \gamma_2 t^2 + g(\mathbf{x}_i)\right] dt} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(y - \mathbf{x}^T\boldsymbol{\beta})^2/(2\sigma^2) + \gamma_1 y + \gamma_2 y^2\right]}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(t - \mathbf{x}^T\boldsymbol{\beta})^2/(2\sigma^2) + \gamma_1 t + \gamma_2 t^2\right] dt}. \end{aligned}$$

Note that

$$\begin{aligned} -\frac{(t - \mathbf{x}^T\boldsymbol{\beta})^2}{2\sigma^2} + \gamma_1 t + \gamma_2 t^2 &= \frac{-t^2 + 2\mathbf{x}^T\boldsymbol{\beta} t - (\mathbf{x}^T\boldsymbol{\beta})^2 + 2\gamma_1\sigma^2 t + 2\gamma_2\sigma^2 t^2}{2\sigma^2} \\ &= \frac{-t^2[1 - 2\gamma_2\sigma^2] + 2t[\mathbf{x}^T\boldsymbol{\beta} t + \gamma_1\sigma^2] - (\mathbf{x}^T\boldsymbol{\beta})^2}{2\sigma^2} \\ &= \frac{-t^2 + 2t[\mathbf{x}^T\boldsymbol{\beta} t + \gamma_1\sigma^2]/[1 - 2\gamma_2\sigma^2] - (\mathbf{x}^T\boldsymbol{\beta})^2/[1 - 2\gamma_2\sigma^2]}{2\sigma^2/[1 - 2\gamma_2\sigma^2]} \\ &= \frac{\left(t - \frac{\mathbf{x}^T\boldsymbol{\beta} t + \gamma_1\sigma^2}{1 - 2\gamma_2\sigma^2}\right)^2 - \frac{(\mathbf{x}^T\boldsymbol{\beta})^2}{1 - 2\gamma_2\sigma^2} + \left(\frac{\mathbf{x}^T\boldsymbol{\beta} t + \gamma_1\sigma^2}{1 - 2\gamma_2\sigma^2}\right)^2}{2\sigma^2/[1 - 2\gamma_2\sigma^2]}. \end{aligned}$$

Thus, the distribution of $Y|Z = 1, \mathbf{x}$ is normal with mean $(\gamma_1\sigma^2 + \mathbf{x}_i^T\boldsymbol{\beta})/(1 - 2\gamma_2\sigma^2)$ and variance $\sigma^2/(1 - 2\gamma_2\sigma^2)$.

11.26 From (11.12), 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

$$\hat{\mathbf{B}} = \frac{1}{\left(\sum_{i\in\mathcal{S}} w_i\right)\left(\sum_{i\in\mathcal{S}} w_i x_i^2\right) - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)^2} \begin{bmatrix} \sum_{i\in\mathcal{S}} w_i x_i^2 & -\sum_{i\in\mathcal{S}} w_i x_i \\ -\sum_{i\in\mathcal{S}} w_i x_i & \sum_{i\in\mathcal{S}} w_i \end{bmatrix} \begin{bmatrix} \sum_{i\in\mathcal{S}} w_i y_i \\ \sum_{i\in\mathcal{S}} w_i x_i y_i \end{bmatrix}.$$

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

$$\begin{aligned}\hat{B}_0 &= \frac{\left(\sum_{i\in\mathcal{S}} w_i x_i^2\right)\left(\sum_{i\in\mathcal{S}} w_i y_i\right) - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)\left(\sum_{i\in\mathcal{S}} w_i x_i y_i\right)}{\left(\sum_{i\in\mathcal{S}} w_i\right)\left(\sum_{i\in\mathcal{S}} w_i x_i^2\right) - \left(\sum_{i\in\mathcal{S}} w_i x_i\right)^2} \\ &= \left(\sum_{i\in\mathcal{S}} w_i\right)^{-1}\left[\sum_{i\in\mathcal{S}} w_i y_i - \hat{B}_1 \sum_{i\in\mathcal{S}} w_i x_i\right].\end{aligned}$$

11.27 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,

$$\begin{aligned}\mathbf{z}_i &= \frac{\partial \hat{\mathbf{B}}}{\partial w_i} \\ &= \left[\frac{\partial}{\partial w_i}\left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1}\right] \sum_{j\in\mathcal{S}} w_j \mathbf{x}_j y_j + \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1}\left[\frac{\partial}{\partial w_i}\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j y_j\right] \\ &= -\left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_i \mathbf{x}_i^T \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \sum_{j\in\mathcal{S}} w_j \mathbf{x}_j y_j + \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_i y_i \\ &= \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \left(-\mathbf{x}_i \mathbf{x}_i^T \hat{\mathbf{B}} + \mathbf{x}_i y_i\right) \\ &= \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_i \left(y_i - \mathbf{x}_i^T \hat{\mathbf{B}}\right).\end{aligned}$$

Then the estimated variance is

$$\begin{aligned}\hat{V}(\hat{\mathbf{B}}) &= \hat{V}\left(\sum_{i\in\mathcal{S}} w_i \mathbf{z}_i\right) \\ &= \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \hat{V}\left(\sum_{i\in\mathcal{S}} w_i \mathbf{q}_i\right) \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1}.\end{aligned}$$

11.28 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,

$$\begin{aligned}\frac{\partial \hat{\mathbf{B}}}{\partial w_i} &= \left[\frac{\partial}{\partial w_i}\left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1}\right] \sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j y_j \\ &\quad + \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \left[\frac{\partial}{\partial w_i} \sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j y_j\right] \\ &= -\left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \frac{1}{\sigma_i^2} \mathbf{x}_i \mathbf{x}_i^T \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j y_j \\ &\quad + \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \frac{1}{\sigma_i^2} \mathbf{x}_i y_i \\ &= \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \left(-\frac{1}{\sigma_i^2} \mathbf{x}_i \mathbf{x}_i^T \hat{\mathbf{B}} + \frac{1}{\sigma_i^2} \mathbf{x}_i y_i\right) \\ &= \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \frac{1}{\sigma_i^2} \mathbf{x}_i \left(y_i - \mathbf{x}_i^T \hat{\mathbf{B}}\right)\end{aligned}$$

$$z_i = \frac{\partial \hat{t}_{y\text{GREG}}}{\partial w_i}$$

$$
\begin{aligned}
&= y_i - \mathbf{x}_i^T\hat{\mathbf{B}} + \left(\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x}\right)^T \frac{\partial\hat{\mathbf{B}}}{\partial w_i} \\
&= y_i - \mathbf{x}_i^T\hat{\mathbf{B}} + \left(\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x}\right)^T \left(\sum_{j\in\mathcal{S}} w_j \frac{1}{\sigma_j^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1} \frac{1}{\sigma_i^2}\mathbf{x}_i\left(y_i - \mathbf{x}_i^T\hat{\mathbf{B}}\right) \\
&= g_i(y_i - \mathbf{x}_i^T\hat{\mathbf{B}}).
\end{aligned}
$$

11.29 The following R code will calculate the values of g_i for *deadtrees*.

```
data(deadtrees)
deadtrees$sampwt = rep(100/25,nrow(deadtrees))
totalvec = c(100,1130)
#lmobj = lm(deadtrees$field~deadtrees$photo,weights=deadtrees$sampwt)
thatx = crossprod(deadtrees$sampwt,deadtrees$photo)
thatxsq = crossprod(deadtrees$sampwt,deadtrees$photo^2)
mat = cbind( c(sum(deadtrees$sampwt), thatx), c(thatx, thatxsq))
deadtrees$gwt = rep(0,nrow(deadtrees))
deadtrees$gwt
##  [1] 0.9535398 1.1084071 0.7212389 1.1858407 1.1858407 0.6438053 1.4955752 1.4181416
    1.3407080
## [10] 0.9535398 1.2632743 1.1084071 0.9535398 0.5663717 1.1084071 0.9535398 0.9535398
    0.8761062
## [19] 0.6438053 1.0309735 0.7212389 0.8761062 1.0309735 0.9535398 0.9535398

for (i in 1:nrow(deadtrees)) {
  deadtrees$gwt[i] = 1 + t( totalvec - mat[,1]) %*% solve(mat) %*% c(1,deadtrees$photo[i
    ])
}
sum(deadtrees$gwt)  # Note that sum of g-weights = 25 because Nhat = N
## [1] 25

sqrt(var(deadtrees$gwt))/mean(deadtrees$gwt)

# Now look at Vhat1, Vhat2
deadtrees$resid = lm(deadtrees$field~deadtrees$photo)$resid
deadtrees$gresid = deadtrees$resid*deadtrees$gwt

dtree<- svydesign(id = ~1, weights=~sampwt, fpc=rep(100,25), data = deadtrees)
# Look at SEs for three totals: y, e_i, g_i e_i
svytotal(~field + resid + gresid,dtree)
##             total     SE
## field  1.1560e+03 52.221
## resid  1.9984e-15 40.798
## gresid 2.3315e-15 41.802

# Fit with survey regression, to see what survey uses
svyglmfit = svyglm(field~photo, design=dtree)
predict(svyglmfit,data.frame(photo=1130),total=100)
```

```
##     link     SE
##  1 1198.9 41.802
```

Note that SAS software will give the same answer as R if you use the VADJUST=none option.

11.31 (a) Let $\mathbf{X}_\mathcal{S}$ be the $n \times p$ matrix whose rows are $\mathbf{x}_i$ for $i \in \mathcal{S}$, and let $\mathbf{W}_\mathcal{S}$ be the $n \times n$ diagonal matrix with diagonal entries $w_i, i \in \mathcal{S}$. Let tr denote the trace of a matrix.

$$\begin{aligned}
\sum_{i\in\mathcal{S}} h(\mathcal{S})_i &= \sum_{i\in\mathcal{S}} w_i \mathbf{x}_i^T \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_i \\
&= \sum_{i\in\mathcal{S}} \sqrt{w_i}\mathbf{x}_i^T \left(\sum_{j\in\mathcal{S}} w_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_i\sqrt{w_i} \\
&= tr\left[\mathbf{W}_\mathcal{S}^{1/2}\mathbf{X}_\mathcal{S}\left(\mathbf{X}_\mathcal{S}^T\mathbf{W}_\mathcal{S}\mathbf{X}_\mathcal{S}\right)^{-1}\mathbf{X}_\mathcal{S}\mathbf{W}_\mathcal{S}^{1/2}\right] \\
&= tr\left[\mathbf{X}_\mathcal{S}\mathbf{W}_\mathcal{S}^{1/2}\mathbf{W}_\mathcal{S}^{1/2}\mathbf{X}_\mathcal{S}\left(\mathbf{X}_\mathcal{S}^T\mathbf{W}_\mathcal{S}\mathbf{X}_\mathcal{S}\right)^{-1}\right] \\
&= tr\left[\mathbf{I}_p\right] \qquad\qquad = p.
\end{aligned}$$

(b) You can calculate the leverage and DFFITS in SAS software with the following code. (You can also use the IML procedure for more efficient matrix calculations, in case you want to calculate these statistics for models with more parameters.)

```
   weight wt;
   model height = finger/clparm covb inverse;
   output out=regpred pred=predicted residual=resid stdp = stdp;
   ods output invxpx = invxpx;
   title 'Regression of anthuneq saving residuals';

data invxpx;
   set invxpx;
   if parameter = "Intercept" then do;
       call symput("M11",Intercept);
       call symput("M12",finger);
   end;
   if parameter = "finger" then call symput("M22",finger);

data leverage;
   set regpred;
   leverage = wt*(&M11 + 2*&M12*finger + &M22*finger*finger);
   dffits = leverage*resid / ( (1-leverage)*stdp);
run;
```

The following table shows the 10 observations with the largest values of DFFITS. Note that observation 97, with finger length = 12.2 cm and height 71 in, is highly influential.

Obs no.	finger	height	wt	resid	predicted	stdp	leverage	dffits
97	12.2	71	174.757	3.55413	67.4459	0.60064	0.15250	1.06471
109	11.0	67	87.379	3.21905	63.7810	0.27775	0.05551	0.68111
193	11.1	67	87.379	2.91364	64.0864	0.25008	0.04623	0.56478
94	12.4	69	174.757	0.94331	68.0567	0.71091	0.21542	0.36432
184	11.3	66	87.379	1.30282	64.6972	0.23278	0.03369	0.19512
56	10.4	65	14.563	3.05151	61.9485	0.55750	0.02553	0.14337
132	10.4	65	14.563	3.05151	61.9485	0.55750	0.02553	0.14337
9	11.6	67	87.379	1.38659	65.6134	0.30627	0.02987	0.13941
63	11.4	66	87.379	0.99741	65.0026	0.24584	0.03042	0.12727
49	11.7	67	87.379	1.08118	65.9188	0.34773	0.03260	0.10478

11.32 See solution to Exercise 11.31. Observation 97, with finger length = 12.2 cm and height 71 in, has the highest value of DFFITS and also has high leverage because of its large weight. For the jackknife replicate with this observation deleted, the estimated equation is

$$\hat{y} = 36.5709460 + 2.4783124x.$$

By contrast, the equation for the model with all of the data is $\hat{y} = 30.1859 + 3.0541x$, so you can see how much influence this point has in determining the regression equation.

This is the reason the jackknife variance is so much higher than the linearization variance in Example 11.6.

11.33 The estimates in (a) and (b) will differ because the straight line model does not describe the data well. A graph of the data indicates a different relationship, overall, in the private colleges than in the public colleges. Adding the extra terms in the model means that a separate model is fit to each stratum and for that case the weighted and unweighted estimates are the same.

11.35 The OLS estimator of $\boldsymbol{\beta}$ is

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

Thus,

$$\begin{aligned} \mathrm{Cov}_M(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Cov}\,(\mathbf{Y})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

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 (11.9) provides an estimate of $V_M(\hat{\beta}_1)$.

11.36 For this model (see Exercise 26)

$$\mathbf{X}_\mathcal{S}^T\mathbf{W}_\mathcal{S}\mathbf{\Sigma}_\mathcal{S}^{-1}\mathbf{X}_\mathcal{S} = \frac{1}{\sigma^2}\begin{bmatrix} \sum_{i\in\mathcal{S}} w_i & \sum_{i\in\mathcal{S}} w_i x_i \\ \sum_{i\in\mathcal{S}} w_i x_i & \sum_{i\in\mathcal{S}} w_i x_i^2 \end{bmatrix}$$

and

$$\mathbf{X}_\mathcal{S}^T\mathbf{W}_\mathcal{S}\mathbf{\Sigma}_\mathcal{S}^{-1}\mathbf{y}_\mathcal{S} = \frac{1}{\sigma^2}\begin{bmatrix} \sum_{i\in\mathcal{S}} w_i y_i \\ \sum_{i\in\mathcal{S}} w_i x_i y_i \end{bmatrix} = \frac{1}{\sigma^2}\frac{\hat{t}_y}{\hat{t}_{xy}}.$$

Using (11.24),

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{N} & \hat{t}_x \\ \hat{t}_x & \hat{t}_{x^2} \end{bmatrix}^{-1}\begin{bmatrix} \hat{t}_y \\ \hat{t}_{xy} \end{bmatrix} = \frac{1}{\hat{N}\hat{t}_{x^2} - (\hat{t}_x)^2}\begin{bmatrix} \hat{t}_{x^2}\hat{t}_y - \hat{t}_x\hat{t}_{xy} \\ -\hat{t}_x\hat{t}_y + \hat{N}\hat{t}_{xy} \end{bmatrix}$$

and

$$\begin{aligned} \hat{t}_{y\mathrm{GREG}} &= \hat{t}_y + [N-\hat{N}, \quad t_x - \hat{t}_x]\hat{\mathbf{B}} \\ &= \hat{t}_y + \frac{1}{\hat{N}\hat{t}_{x^2} - (\hat{t}_x)^2}[(N-\hat{N})(\hat{t}_{x^2}\hat{t}_y - \hat{t}_x\hat{t}_{xy}) + (t_x-\hat{t}_x)(-\hat{t}_x\hat{t}_y + \hat{N}\hat{t}_{xy})]. \end{aligned}$$

If $y = x$,

$$\begin{aligned}\hat{t}_{x\text{GREG}} &= \hat{t}_x + \frac{1}{\hat{N}\hat{t}_{x^2} - (\hat{t}_x)^2}[0 + (t_x - \hat{t}_x)(-\hat{t}_x^2 + \hat{N}\hat{t}_{x^2})] \\ &= t_x.\end{aligned}$$

11.37 (a) Define indicator variables $x_{ih} = 1$ if population unit i is in stratum h, and 0 otherwise. With these x variables, stratified random sampling satisfies the balancing equation (11.30).

(b) A stratified sample can be selected so that the sample percentages in each gender/race/education group equals the population percentages. But this requires you to form $2 \times 5 \times 5 = 50$ strata, and some of those strata may be small. The balanced sample does not require the equality of the percentages within each combination of the variables, but requires it separately for each characteristics. Thus, the balanced sample would, in a typical population, have about half men and half women. But it isn't required to have half men and half women in each education or race group.

In addition, stratified sampling requires that the stratification variables be categorical. Balanced sampling places no such restriction.

11.38 *Calibration.* The method of Lagrange multipliers considers the function

$$h(\mathbf{w}, \tilde{\mathbf{w}}) = \sum_{i \in \mathcal{S}} \frac{(w_i - \tilde{w}_i)^2}{w_i} + \boldsymbol{\lambda}^T \left(\sum_{i \in \mathcal{S}} \tilde{w}_i \mathbf{x}_i - \mathbf{t}_\mathbf{x} \right).$$

We now take the derivative of g with respect to each $\tilde{w}_i$ and with respect to λ, and set equal to 0.

$$\frac{\partial h(\mathbf{w}, \tilde{\mathbf{w}})}{\partial \tilde{w}_i} = -2\frac{w_i - \tilde{w}_i}{w_i} + \boldsymbol{\lambda}^T \mathbf{x}_i = 0 \tag{11.1}$$

$$\frac{\partial h(\mathbf{w}, \tilde{\mathbf{w}})}{\partial \boldsymbol{\lambda}} = \left(\sum_{i \in \mathcal{S}} \tilde{w}_i \mathbf{x}_i - \mathbf{t}_\mathbf{x} \right) = 0. \tag{11.2}$$

Multiplying each side of (SM11.1) by $w_i \mathbf{x}_i^T$ and summing, we have

$$-2 \sum_{i \in \mathcal{S}} w_i \mathbf{x}_i^T + 2 \sum_{i \in \mathcal{S}} \tilde{w}_i \mathbf{x}_i^T + \boldsymbol{\lambda}^T \sum_{i \in \mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T = 0$$

Under the calibration constraint in (SM11.2), this means

$$-2\mathbf{t}_\mathbf{x} + 2\hat{\mathbf{t}}_\mathbf{x} + \left(\sum_{i \in \mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T \right) \boldsymbol{\lambda} = 0$$

This gives

$$\boldsymbol{\lambda} = 2\left(\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1}\left(\hat{\mathbf{t}}_{\mathbf{x}} - \mathbf{t}_{\mathbf{x}}\right).$$

Now, substituting the value of $\boldsymbol{\lambda}$ into (SM11.1) gives

$$\begin{aligned}\tilde{w}_i &= w_i\left[1 + \frac{1}{2}\boldsymbol{\lambda}^T\mathbf{x}_i\right] \\ &= w_i\left[1 + \left(\hat{\mathbf{t}}_{\mathbf{x}} - \mathbf{t}_{\mathbf{x}}\right)^T\left(\sum_{i\in\mathcal{S}} w_i\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\mathbf{x}_i\right].\end{aligned}$$

11.39 (a) Using the indicator variables for the weighting classes, the auxiliary H-vector for unit i is $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iH}]^T$. Note that there is no intercept component for this vector, and $\mathbf{x}_i$ contains the entry 1 in position c if unit i is in poststratum h and zeroes elsewhere. Then, $\sum_{j\in\mathcal{R}} w_j\mathbf{x}_j\mathbf{x}_j^T$ s the $H \times H$ diagonal matrix where the hth diagonal entry is $\sum_{j\in\mathcal{R}} w_j x_{jh}$. Then, for unit i in poststratum h,

$$\left(\mathbf{t}_{\mathbf{x}} - \sum_{j\in\mathcal{R}} w_j\mathbf{x}_j\right)^T\left(\sum_{j\in\mathcal{R}} w_j\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\mathbf{x}_i = \frac{N_h - \sum_{j\in\mathcal{R}} w_j x_{jh}}{\sum_{j\in\mathcal{R}} w_j x_{jh}}$$

and

$$\begin{aligned}w_i^* &= w_i\left[1 + \frac{N_h - \sum_{j\in\mathcal{R}} w_j x_{jh}}{\sum_{j\in\mathcal{R}} w_j x_{jh}}\right] \\ &= w_i\frac{N_h}{\sum_{j\in\mathcal{R}} w_j x_{jh}}.\end{aligned}$$

(b) $\sum_{j=1}^N \phi_j\mathbf{x}_j\mathbf{x}_j^T$ is the $H \times H$ diagonal matrix where the hth diagonal entry is $\sum_{j=1}^N \phi_j x_{jh}$, the sum of the response propensities (for everyone in the population) in poststratum h. Then,

$$\begin{aligned}\text{Bias} &\approx -\sum_{i=1}^N(1-\phi_i)\left[y_i - \mathbf{x}_i^T\left(\sum_{j=1}^N \phi_j\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\sum_{j=1}^N \phi_j\mathbf{x}_j y_j\right] \\ &= -\sum_{i=1}^N\sum_{h=1}^H(1-\phi_i)x_{ih}\left[y_i - \frac{\sum_{j=1}^N \phi_j x_{jh}y_j}{\sum_{j=1}^N \phi_j x_{jh}}\right] \\ &= -\sum_{h=1}^H\sum_{i=1}^N x_{ih}(1-\phi_i)\left[y_i - \frac{\sum_{j=1}^N \phi_j x_{jh}y_j}{\sum_{j=1}^N \phi_j x_{jh}}\right]\end{aligned}$$

Note that the expression for the bias depends only on the response propensities, not on the sampling design.

11.40

Age Group	Gender	Population Size	Sample Size	g_i (Poststratification)	g_i (Model)
25 or under	Female	50	15	0.1667	−0.170
25 or under	Male	250	26	0.4808	0.675
Over 25	Female	100	5	1.0000	2.01
Over 25	Male	900	14	3.2143	2.85

11.41 We can prove this result using successive conditioning.

For this to be more efficient, the gains from poststratification must outweigh the extra variance from not knowing the control totals. The second term in (11.35) can have the same order of magnitude as the first term (and even be larger if the auxiliary survey has high variability).

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

$$\begin{aligned}
\hat{V}\left(\hat{t}_{\text{str}}^{(2)}\right) &= N(N-1)\sum_{h=1}^{H}\left(\frac{n_h-1}{n-1} - \frac{m_h-1}{N-1}\right)\frac{n_h}{n}\frac{s_h^{2(2)}}{m_h} \\
&\qquad + \frac{N^2}{n-1}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{n_h}{n}(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2 \\
&= 100000(99999)0.000912108 + \frac{100000^2}{999}0.109282588 \\
&= 10214904.
\end{aligned}$$

Thus $\text{SE}(\hat{t}_{\text{str}}^{(2)}) = 3196$.

We can obtain the with-replacement variance using R:

```
> svytotal(~diabexam,d1201)
         total     SE
diabexam 48310 3228.5
```

12.2 Answers will vary.

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 this data set was reconstructed from information in the paper and it is possible that the results from these data do not reflect the distribution of shorebirds in the region.

12.4 (a) The phase II 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

$$\hat{\bar{y}}_{\text{str}}^{(2)} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(2)} = 0.3030 + 0.1078 + 0.1463 = 0.5570.$$

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 II 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:

		Case?		
Design effect		No	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 χ_1^2 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.

This can also be calculated in R. First, create a data frame containing 1558 rows with variables y and *male*. Then we have

```
d1205 <- twophase(id=list(~1,~1), weights=list(~p1weight, ~p2weight),
        strata=list(NULL,~stratum), subset=~inp2, data=exer1204df, method="simple")

svychisq(~male+y,d1205)
##
##         Pearson's X^2: Rao & Scott adjustment

## data:  svychisq(~male + y, d1205)
## F = 7.1493, ndf = 1, ddf = 247, p-value = 0.007999
```

12.6 The estimated total is 5267, with 95% CI [2409, 8125].

12.7 The estimated total is 9316, with 95% CI [5733.4, 12899.0].

12.8 The estimated total is 7306, with 95% CI [4187, 10424].

12.9 (a) The CV of *p1weight* is 0.39; the CV of *p2weight* is 1.16.

(b) Because the subsampling was done within strata, we can view this as an unequal probability sample within each stratum. Here are results from R (using with-replacement variance).

```
d1209 <- svydesign(id=~1,strata=~stratum,weights=~p2weight,data=mysteries)
svytotal(~victims,d1209,na.rm=T)
##          total     SE
## victims 2327.5 485.95
```

(c) Books written by women have an average of 1.96 victims; those by men an average of 4.10 victims. The difference is 2.15, with 95% CI [0.47, 3.83].

(d) The table from R is given below. The Rao–Scott second-order F statistic is 0.38, with p-value > 0.5. There is no evidence of an association.

```
             bfirearm
authorgender         0         1
           F  78.33333  87.58333
           M 175.00000 314.08333
```

12.10 We estimate W_h by n_h/n, and estimate S_h^2 by $s_h^{2(2)}$. Then, using (12.18), we have

Stratum	$\hat{W}_h$	$\hat{S}_h^2$	$\hat{W}_h\hat{S}_h^2$	ν_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 $\nu_1 = \nu_2 = \nu_3$.

12.11 From Property 5 in Table A.2,

$$V(\hat{t}_y^{(2)}) = V(\hat{t}_y^{(1)}) + E(V[\hat{t}_y^{(2)} \mid \mathbf{Z}]).$$

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,

$$P(D_i = 1 \mid Z_i = 1) = n^{(2)}\Big/n^{(1)},$$

$$P(D_iD_j = 1 \mid Z_iZ_j = 1) = \frac{n^{(2)}(n^{(2)}-1)}{n^{(1)}(n^{(1)}-1)} \text{ for } j \neq i,$$

$$w_i^{(1)} = \frac{N}{n^{(1)}},$$

and

$$w_i^{(2)} = \frac{n^{(1)}}{n^{(2)}}Z_i.$$

In addition,

$$P(Z_i = 1) = \frac{n^{(1)}}{N}$$

and

$$P(Z_iZ_j = 1) = \frac{n^{(1)}(n^{(1)}-1)}{N(N-1)}.$$

Thus, using (12.1) to write $\hat{t}_y^{(2)}$,

$$V[\hat{t}_y^{(2)} \mid \mathbf{Z}] = V\left[\sum_{i=1}^{N} Z_iD_i\frac{N}{n^{(1)}}\frac{n^{(1)}}{n^{(2)}}y_i \mid \mathbf{Z}\right]$$

$$= \left(\frac{N}{n^{(2)}}\right)^2 E\Big[\sum_{i=1}^N\sum_{k=1}^N Z_i Z_k D_i D_k y_i y_k \mid \mathbf{Z}\Big] - [\hat{t}_y^{(1)}]^2$$
$$= \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^N Z_i y_i^2 \frac{n^{(2)}}{n^{(1)}} + \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^N \sum_{j=1,j\neq i}^N Z_i Z_j y_i y_j \frac{n^{(2)}(n^{(2)}-1)}{n^{(1)}(n^{(1)}-1)} - [\hat{t}_y^{(1)}]^2.$$

Also,

$$\begin{aligned} E(V[\hat{t}_y^{(2)} \mid \mathbf{Z}]) &= \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^N \frac{n^{(1)}}{N} y_i^2 \frac{n^{(2)}}{n^{(1)}} \\ &\quad + \left(\frac{N}{n^{(2)}}\right)^2 \sum_{i=1}^N \sum_{j=1,j\neq i}^N \frac{n^{(1)}(n^{(1)}-1)}{N(N-1)} y_i y_j \frac{n^{(2)}(n^{(2)}-1)}{n^{(1)}(n^{(1)}-1)} \\ &\quad - V[\hat{t}_y^{(1)}] - (E[\hat{t}_y^{(1)}])^2 \\ &= \frac{N}{n^{(2)}} \sum_{i=1}^N y_i^2 + \frac{N(n^{(2)}-1)}{n^{(2)}(N-1)} \sum_{i=1}^N \sum_{j=1,j\neq i}^N y_i y_j \\ &\quad - V[\hat{t}y^{(1)}] - t_y^2 \\ &= N^2\left(1 - \frac{n^{(2)}}{N}\right)\frac{S_y^2}{n^{(2)}} - V[\hat{t}_y^{(1)}]. \end{aligned}$$

Thus,

$$V[\hat{t}_y^{(2)}] = N^2\left(1 - \frac{n^{(2)}}{N}\right)\frac{S_y^2}{n^{(2)}}.$$

12.12 *Estimated variance in two-phase sampling with stratification.* Conditioning on the phase I units,

$$E[\hat{V}(\hat{t}_{\text{str}}^{(2)})|\mathbf{Z}] = N(N-1)\sum_{h=1}^H \left(\frac{n_h - 1}{n-1} - \frac{m_h - 1}{N-1}\right)\frac{n_h}{n} E\left[\frac{s_h^{2(2)}}{m_h}\Big|\mathbf{Z}\right] \tag{12.1}$$
$$+ \frac{N^2}{n-1}\left(1 - \frac{n}{N}\right)\sum_{h=1}^H E\Big[\frac{n_h}{n}(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2\Big|\mathbf{Z}\Big]. \tag{12.2}$$

Now

$$E\left[\frac{s_h^{2(2)}}{m_h}\Big|\mathbf{Z}\right] = \frac{s_h^{2(1)}}{m_h}$$

and

$$\sum_{h=1}^H E\left[\frac{n_h}{n}(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{str}}^{(2)})^2\Big|\mathbf{Z}\right] = E\left[\sum_{h=1}^H \frac{n_h}{n}(\bar{y}_h^{(2)})^2 - (\hat{\bar{y}}_{\text{str}}^{(2)})^2\Big|\mathbf{Z}\right]$$
$$= \sum_{h=1}^H \frac{n_h}{n}\left[\frac{s_h^{2(1)}}{m_h}\left(1 - \frac{m_h}{n_h}\right) + (\bar{y}_h^{(1)})^2\right]$$

$$
\begin{aligned}
&- \sum_{h=1}^{H} \left(\frac{n_h}{n}\right)^2 \left(1 - \frac{m_h}{n_h}\right) \frac{s_h^{2(1)}}{m_h} - \left[\sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(1)}\right]^2 \\
&= \sum_{h=1}^{H} \frac{n_h}{n}\left(1 - \frac{n_h}{n}\right)\left(1 - \frac{m_h}{n_h}\right) \frac{s_h^{2(1)}}{m_h} + \frac{1}{n}\left[(n-1)s_y^{2(1)} - \sum_{h=1}^{H}(n_h - 1)s_h^{2(1)}\right];
\end{aligned}
$$

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 (SM12.2), 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 $E[s_y^{2(1)}] = S_y^2$ and $\frac{s_h^{2(1)}}{m_h}\left(1 - \frac{m_h}{n_h}\right) \approx V(\bar{y}_h^{(2)})$, the estimator is approximately unbiased.

12.13 (a) Equation (A.9) implies these results.

(b) From the solution to Exercise 12.11,

$$
\begin{aligned}
V(\hat{t}_{yr}^{(2)}) &= V[\hat{t}_y^{(1)}] + E[V(\hat{t}_d^{(2)} \mid \mathbf{Z})], \\
V[\hat{t}_y^{(1)}] &= N^2\left(1 - \frac{n^{(1)}}{N}\right)\frac{S_y^2}{n^{(1)}},
\end{aligned}
$$

and

$$
\begin{aligned}
E[V(\hat{t}_d^{(2)} \mid \mathbf{Z})] &= N^2\left(1 - \frac{n^{(2)}}{N}\right)\frac{S_d^2}{n^{(2)}} - V[\hat{t}_d^{(1)}] \\
&= N^2\left(1 - \frac{n^{(2)}}{N}\right)\frac{S_d^2}{n^{(2)}} - N^2\left(1 - \frac{n^{(1)}}{N}\right)\frac{S_d^2}{n^{(1)}} \\
&= NS_d^2\left[\frac{N - n^{(2)}}{n^{(2)}} - \frac{N - n^{(1)}}{n^{(1)}}\right]
\end{aligned}
$$

$$= N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}}.$$

(c) Follows because s_y^2 and s_e^2 estimate S_y^2 and S_d^2, respectively.

12.14 (a) 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,

$$\begin{aligned}
V(\hat{t}_{yr}^{(2)}) &\approx N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{S_y^2}{n^{(1)}}+N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}} \\
&= N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{2BS_{xd}+B^2S_x^2+S_d^2}{n^{(1)}}+N^2\left(1-\frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}} \\
&= N^2\left(1-\frac{n^{(1)}}{N}\right)\frac{2BS_{xd}+B^2S_x^2}{n^{(1)}}+N^2\left(1-\frac{n^{(2)}}{N}\right)\frac{S_d^2}{n^{(2)}}.
\end{aligned}$$

12.15 *Demnati-Rao* (a)

$$z_i^{(1)} = \frac{\partial \hat{t}_{yr}^{(2)}}{\partial w_i^{(1)}} = x_i\frac{\hat{t}_y^{(2)}}{\hat{t}_x^{(2)}}$$

and

$$z_i^{(2)} = \frac{\partial \hat{t}_{yr}^{(2)}}{\partial w_i} = \hat{t}_x^{(1)}\left[\frac{y_i}{\hat{t}_x^{(2)}}-\frac{x_i}{\hat{t}_x^{(2)}}\frac{\hat{t}_y^{(2)}}{\hat{t}_x^{(2)}}\right] = \frac{\hat{t}_x^{(1)}}{\hat{t}_x^{(2)}}\left[y_i-x_i\hat{t}_x^{(2)}\hat{B}^{(2)}\right]$$

Thus,

$$\begin{aligned}
\hat{V}_{\mathrm{DR}}(\hat{t}_{yr}^{(2)}) &= \hat{V}\left(\sum_{i\in\mathcal{S}^{(1)}}w_i^{(1)}z_i^{(1)}+\sum_{i\in\mathcal{S}^{(2)}}w_iz_i^{(2)}\right) \\
&= \hat{V}\left(\sum_{i\in\mathcal{S}^{(1)}}w_i^{(1)}x_i\hat{B}^{(2)}+\sum_{i\in\mathcal{S}^{(2)}}w_i\frac{\hat{t}_x^{(1)}}{\hat{t}_x^{(2)}}\left[y_i-x_i\hat{t}_x^{(2)}\hat{B}^{(2)}\right]\right)
\end{aligned}$$

12.17 Note that

$$\begin{aligned} E[\hat{V}_{HT}(\hat{t}_y^{(1)}) \mid \mathbf{Z}] &= E\left[\sum_{i\in\mathcal{S}^{(2)}}\sum_{k\in\mathcal{S}^{(2)}} \frac{\pi_{ik}^{(1)}-\pi_i^{(1)}\pi_k^{(1)}}{\pi_{ik}^{(1)}\pi_{ik}^{(2)}}\frac{y_i}{\pi_i^{(1)}}\frac{y_k}{\pi_k^{(1)}} \mid \mathbf{Z}\right] \\ &= E\left[\sum_{i\in\mathcal{S}^{(1)}}\sum_{k\in\mathcal{S}^{(1)}} D_iD_k\frac{\pi_{ik}^{(1)}-\pi_i^{(1)}\pi_k^{(1)}}{\pi_{ik}^{(1)}\pi_{ik}^{(2)}}\frac{y_i}{\pi_i^{(1)}}\frac{y_k}{\pi_k^{(1)}} \mid \mathbf{Z}\right] \\ &= E\left[\sum_{i\in\mathcal{S}^{(1)}}\sum_{k\in\mathcal{S}^{(1)}} \frac{\pi_{ik}^{(1)}-\pi_i^{(1)}\pi_k^{(1)}}{\pi_{ik}^{(1)}}\frac{y_i}{\pi_i^{(1)}}\frac{y_k}{\pi_k^{(1)}} \mid \mathbf{Z}\right] \\ &= \hat{V}_{HT}^{(1)}(\hat{t}_y^{(1)}) \end{aligned}$$

12.18 (a) To avoid subscripts, let $n = n^{(1)}$, $s_h^{2(1)} = s_h^2$, and $c = c^{(1)}$. Since $m_h = \nu_h n_h$ and ν_h is known,

$$E[m_h] = \nu_h E[n_h] = \nu_h E\Big[\sum_{i=1}^N Z_i x_{ih}\Big] = \nu_h \sum_{i=1}^N \frac{n}{N} x_{ih} = n\nu_h W_h.$$

Thus,

$$E[C] = cn + n\sum_{h=1}^H c_h\nu_h W_h.$$

(b) This proof is given in Theorem 1 of Rao (1973) and on page 328 of Cochran (1977). From (12.6),

$$V\left(y_{\text{str}}^{(2)}\right) = \left(1-\frac{n}{N}\right)\frac{S_y^2}{n} + E\left[\sum_{h=1}^H \left(\frac{n_h}{n}\right)^2\left(1-\frac{m_h}{n_h}\right)\frac{s_h^2}{m_h}\right].$$

When $m_h = \nu n_h$, the second term is

$$E\left[\sum_{h=1}^H \left(\frac{n_h}{n}\right)^2 (1-\nu_h)\frac{s_h^2}{\nu_h n_h}\right] = \sum_{h=1}^H \left(\frac{1}{\nu_h}-1\right)\frac{1}{n}E\left[n_h s_h^2\right].$$

But

$$\begin{aligned} E\left[n_h s_h^2\right] &= E\left[\frac{n_h}{n_h-1}\sum_{i=1}^N x_{ih}Z_i(y_i-\bar{y}_h)^2\right] \\ &= E\left[\frac{n_h}{n_h-1}\left(\sum_{i=1}^N x_{ih}Z_i y_i^2 - n_h\bar{y}_h^2\right)\right] \end{aligned}$$

$$
\begin{aligned}
V\left(\hat{t}_{\text{str}}^{(2)}\right) &= V\left(E\left[\hat{t}_{\text{str}}^{(2)} \mid \mathbf{Z}\right]\right) + E\left(V\left[\hat{t}_{\text{str}}^{(2)} \mid \mathbf{Z}\right]\right) \\
&= V\left(\hat{t}_{y}^{(1)}\right) + N^2 E\left(V\left[\sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(2)} \,\middle|\, \mathbf{Z}\right]\right) \\
&= N^2\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n} + N^2 E\left[\sum_{h=1}^{H}\left(\frac{n_h}{n}\right)^2\left(1 - \frac{m_h}{n_h}\right)\frac{s_h^{2(1)}}{m_h}\right]. \qquad (12.3)
\end{aligned}
$$

$$
V\left(\hat{\bar{y}}_{\text{str}}^{(2)}\right) = S_y^2\left(\frac{1}{n^{(1)}} - \frac{1}{N}\right) + \frac{1}{n^{(1)}}\sum_{h=1}^{H} W_h S_h^2\left(\frac{1}{\nu_h} - 1\right).
$$

(c) Note that if $S^2 <= \sum_{j=1}^{H} W_j S_j^2$, there would be no reason to use two-phase sampling, since there are no expected efficiency gains.

Using the cost constraint, set

$$
n = \frac{E[C]}{c + \sum_{h=1}^{H} c_h \nu_h W_h}.
$$

Then

$$
V(\hat{\bar{y}}_{\text{str}}^{(2)}) = S^2\left[\frac{c + \sum_{h=1}^{H} c_h \nu_h W_h}{E[C]} - \frac{1}{N}\right] + \frac{c + \sum_{h=1}^{H} c_h \nu_h W_h}{E[C]}\sum_{h=1}^{H} W_h S_h^2\left(\frac{1}{\nu_h} - 1\right)
$$

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

$$
S^2 c_k \nu_k W_k + c_k \nu_k W_k \sum_{h=1}^{H} W_h S_h^2\left(\frac{1}{\nu_h} - 1\right) - \frac{W_k S_k^2}{\nu_k}\left(c + \sum_{h=1}^{H} c_h \nu_h W_h\right) = 0 \qquad (12.4)
$$

for $k = 1, \ldots, H$. Summing over h gives

$$
0 \;=\; \sum_{k=1}^{H} c_k \nu_k W_k\left[S^2 + \sum_{h=1}^{H} W_h S_h^2\left(\frac{1}{\nu_h} - 1\right)\right] - \sum_{k=1}^{H} \frac{W_k S_k^2}{\nu_k}\left(c + \sum_{h=1}^{H} c_h \nu_h W_h\right)
$$

$$= \sum_{k=1}^{H} c_k \nu_k W_k \Big[S^2 - \sum_{h=1}^{H} W_h S_h^2 \Big] - c \sum_{k=1}^{H} \frac{W_k S_k^2}{\nu_k}$$

and

$$\sum_{h=1}^{H} \frac{W_h S_h^2}{\nu_h} = \frac{1}{c} \Big(S^2 - \sum_{h=1}^{H} W_h S_h^2 \Big) \sum_{k=1}^{H} c_k \nu_k W_k.$$

Substituting this expression into (SM12.4), we have

$$\begin{aligned} 0 &= c_k \nu_k W_k \Big[S^2 - \sum_{h=1}^{H} W_h S_h^2 + \frac{1}{c} \Big(S^2 - \sum_{h=1}^{H} W_h S_h^2 \Big) \sum_{h=1}^{H} c_h \nu_h W_h \Big] - \frac{W_k S_k^2}{\nu_k} \Big(c + \sum_{h=1}^{H} c_h \nu_h W_h \Big) \\ &= c_k \nu_k W_k \left(S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) \frac{1}{c} \left(c + \sum_{h=1}^{H} c_h \nu_h W_h \right) - \frac{W_k S_k^2}{\nu_k} \left(c + \sum_{h=1}^{H} c_h \nu_h W_h \right) \\ &= \left[\frac{c_k}{c} \nu_k W_k \left(S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) - \frac{W_k S_k^2}{\nu_k} \right] \left(c + \sum_{h=1}^{H} c_h \nu_h W_h \right). \end{aligned}$$

The second factor is always positive, so

$$\frac{c_k \nu_k^2}{c} \Big(S^2 - \sum_{h=1}^{H} W_h S_h^2 \Big) = S_k^2,$$

and, consequently,

$$\nu_k = \sqrt{\frac{c S_k^2}{c_k (S^2 - \sum_{h=1}^{H} W_h S_h^2)}}.$$

12.19 Let $A = S_y^2 - \sum_{h=1}^{H} W_h S_h^2$. Then, from, (12.18),

$$\nu_{h,opt} = \sqrt{\frac{c^{(1)} S_h^2}{c_h \left(S^2 - \sum_{j=1}^{H} W_j S_j^2 \right)}} = \sqrt{\frac{c^{(1)} S_h^2}{c_h A}}$$

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,

$$V_{opt}(\hat{\bar{y}}_{\text{str}}^{(2)}) = \frac{S_y^2}{n_{opt}^{(1)}} - \frac{S_y^2}{N} + \frac{1}{n_{opt}^{(1)}} \sum_{h=1}^{H} W_h S_h^2 \left(\frac{1}{\nu_{h,opt}} - 1 \right)$$

$$
\begin{aligned}
&= \frac{S_y^2}{n_{opt}^{(1)}} - \frac{S_y^2}{N} + \frac{1}{n_{opt}^{(1)}}\left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\sqrt{\frac{A}{c^{(1)}}} - \sum_{h=1}^{H} W_h S_h^2\right) \\
&= \frac{S_y^2}{n_{opt}^{(1)}} - \frac{S_y^2}{N} + \frac{1}{n_{opt}^{(1)}}\left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\sqrt{\frac{A}{c^{(1)}}} + A - S_y^2\right) \\
&= \frac{1}{n_{opt}^{(1)}}\left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\sqrt{\frac{A}{c^{(1)}}} + A\right) - \frac{S_y^2}{N} \\
&= \frac{1}{C^*}\left(c^{(1)} + \sum_{h=1}^{H} W_h S_h \sqrt{c_h}\sqrt{\frac{c^{(1)}}{A}}\right)\left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\sqrt{\frac{A}{c^{(1)}}} + A\right) - \frac{S_y^2}{N} \\
&= \frac{1}{C^*}\left[\sqrt{c^{(1)}A}\sum_{h=1}^{H} W_h S_h \sqrt{c_h} + \left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\right)^2\right. \\
&\qquad \left. + Ac^{(1)} + \sqrt{c^{(1)}A}\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\right] - \frac{S_y^2}{N} \\
&= \frac{1}{C^*}\left[\sum_{h=1}^{H} W_h S_h \sqrt{c_h} + \sqrt{c^{(1)}}\sqrt{S^2 - \sum_{h=1}^{H} W_h S_h^2}\right]^2 - \frac{S_y^2}{N}.
\end{aligned}
$$

12.20 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 λ 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.

12.21 (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 (12.20),

$$\frac{V_{opt}(\hat{p}_{\text{str}}^{(2)})}{V_{SRS}(\hat{p})} \approx \left[\sum_{h=1}^{2} W_h \frac{S_h}{S_y} + \sqrt{\frac{c^{(1)}}{c^{(2)}}}\sqrt{\frac{S_y^2 - \sum_{h=1}^{2} W_h S_h^2}{S_y^2}}\right]^2.$$

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 $c^{(1)}/c^{(2)}$				
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

12.22 Suppose that $\mathcal{S}$ is a subset of m units from $\mathcal{U}$, 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,

$$\begin{aligned}
&P(F_i = 1, i \in \mathcal{S}) \\
&= \frac{\binom{m_1}{m_1}\binom{N_1 - m_1}{n_1 - m_1}\binom{m_2}{m_2}\binom{N_2 - m_2}{n_2 - m_2}\cdots\binom{m_H}{m_H}\binom{N_H - m_H}{n_H - m_H}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_H}{n_H}}
\end{aligned}$$

$$= \frac{\text{total number of stratified samples containing } \mathcal{S}}{\text{number of possible stratified samples}}$$

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 the Size of a Population

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)
## $cell
## (Intercept)
##         570

## $N
## (Intercept)
##       1250

## $CI_cell
## [1] 436.1055 741.7603

## $CI_N
## [1] 1116.105 1421.760
```

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.

(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].

13.4 Estimate is 16307, with 95% CI [15901, 16742]. You are assuming that participating in the first survey is independent of participating in the second (among those who are survey participants) and that all observations are independent.

13.5 Here are results using the capture-recapture method (Lincoln-Petersen estimator). The $\hat{N}$ column gives the estimated number of accidents for each country. These CIs were calculated using the normal approximation. Other estimates can be used for N and for the CIs.

Country	Sea-Web^{TM}	Flag-State	Both	$\hat{N}$	LCL	UCL
Canada	608	722	454	967	913	1021
Denmark	189	220	45	924	683	1165
Netherlands	304	342	71	1464	1161	1768
Norway	529	596	203	1553	1380	1727
Sweden	109	333	86	422	345	499
United Kingdom	401	1428	229	2501	2204	2797
United States	632	2362	135	11058	9246	12869

13.6 The following table gives the partial contingency table. LI stands for "coded as legal intervention."

	LI	Not LI	Total
In *The Guardian's* list	487	599	1,086
Not in *The Guardian's* list	36		
Total	523		

The total number of deaths is estimated by $(1{,}086)(523)/487 = 1{,}166$. The sample sizes and overlap

are large enough that the normal approximation CI works well. The estimated variance is

$$\hat{V}(\hat{N}) = \frac{(1086)^2(523)(523-487)}{487^3} = 192.3$$

and an approximate 95% CI is

$$1166 \pm 1.96\sqrt{192.3} = [1139, 1193]\,.$$

A bootstrapped CI is similar. Based on 3,000 bootstrap replicates, I calculated a 95% CI as [1143, 1196].

This estimate requires the assumption that presence on *The Guardian's* list is independent of presence on the CDC list—that is, the probability of a death appearing in one list is unrelated to its probability of appearing in the other list. But it may be that some types of deaths are more likely to be hidden from both lists. Both data sets rely heavily on the reporting by police departments, and the news stories used in part to construct *The Guardian's* list are often written in response to information released by a police department. If a law enforcement agency does not release information that a death is attributable to agency actions, then the case is unlikely to appear on either list. Thus, the estimate of 1,166 deaths by law enforcement may be too small.

A note: The authors were actually able to match only 991 of the names on *The Guardian's* list with names on the National Death Index (NDI), finding that 444 of those 991 deaths had been coded as due to legal intervention (LI). They assumed that the 95 unmatched names were classified as LI at the same rate as matched deaths, giving a total of 444 + 95*(444/991) = 487 names on both lists. Thus an additional assumption is that the unmatched names are equally likely to be coded as LI as the matched names.

13.7 (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)
```

13.8 The model with all two-factor interactions has $G^2 = 3.3$, with 4 df. Comparing to a χ^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].

13.9 (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)
$cell
(Intercept)
   128.1224

$N
(Intercept)
   336.1224

$CI_cell
[1]  80.0101 199.6771

$CI_N
[1] 288.0101 407.6771

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 elistdlisttlist@2;
      /* Model with all 2-way intrxns */
run;
```

13.10 (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 (13.3) 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.

Note that there appeared to be a large decrease in cases after 1980.

13.11

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].

13.12 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.

13.13 (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].

13.14 (a) All areas are sampled in the high-density strata, so the sampling rate is 1 for each. Since all are sampled with certainty, it makes no difference whether they are treated as one stratum, or six strata, or as many strata as areas.

(b) 70.4%, with 95% CI [56.4, 82.0]. A very wide CI!

(c) 3000*54/38 = 4263. Because a census is taken in the high-density areas, $\hat{t}_y$ has no sampling variance so

$$V\left[\frac{\hat{t}_y}{\hat{p}}\right] = t_y^2 V\left[\frac{1}{\hat{p}}\right] \approx t_y^2 \frac{1-p}{n_2 p^3}.$$

Substituting estimates, we have

$$\begin{aligned} t_y^2 \hat{V}\left[\frac{1}{\hat{p}}\right] &= (3000)^2 \frac{1-38/54}{54(38/54)^3} \\ &= (3000)^2 * (0.01574574) = 141712. \end{aligned}$$

If we assume a symmetric distribution, this gives an approximate 95% CI of [3508, 5018]. Note that we could also use the bootstrap here to construct an asymmetric CI.

(d) The estimate in (c) is computed assuming that the count missed about 30% of the unsheltered homeless persons. But that estimate is quite imprecise. It's also possible that the decoys behaved differently than the "real" unsheltered persons who were undetected. The estimate is computed assuming that the proportion of uncounted "real" unsheltered persons is the same as the proportion of uncounted decoys. Also, that the decoys are not deliberately placed in easily detectable areas. Note that in 2019, New York City estimated that 90% of the decoys in the shadow count were found.

13.15 Letting $p = m/n_2$, we have

$$\begin{aligned} V\left[\frac{1}{\hat{p}}\right] &\approx V\left[\frac{1}{p^2}(\hat{p}-p)\right] \\ &= \frac{1}{p^4}\frac{p(1-p)}{n_2} \\ &= \frac{1-p}{n_2 p^3}. \end{aligned}$$

Thus,

$$V\left[\hat{N}\right] = V\left[\frac{n_1}{\hat{p}}\right] \approx \frac{n_1^2(1-p)}{n_2 p^3}.$$

13.16 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.

13.17 Note that

$$\mathcal{L}(N-1|n_1,n_2) \;=\; \frac{\binom{n_1}{m}\binom{N-1-n_1}{n_2-m}}{\binom{N-1}{n_2}}$$

$$
\begin{aligned}
&= \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}\dfrac{N-n_1-(n_2-m)}{N-n_1}}{\binom{N}{n_2}\dfrac{N-n_2}{N}} \\
&= \mathcal{L}(N|n_1,n_2)\frac{N-n_1-n_2+m}{N-n_1}\frac{N}{N-n_2}.
\end{aligned}
$$

Thus, if $N > n_1$ and $N > n_2$,

$$
\begin{aligned}
\mathcal{L}(N) \geq \mathcal{L}(N-1) \qquad &\text{iff} \quad \frac{N-n_1-n_2+m}{N-n_1}\frac{N}{N-n_2} \leq 1 \\
&\text{iff} \quad mN \leq n_1 n_2.
\end{aligned}
$$

Take $\hat{N}$ to be the integer part of $n_1 n_2/m$. Then for any integer $k \leq \hat{N}$,

$$
mk \leq m\frac{n_1 n_2}{m} = n_1 n_2,
$$

so $\mathcal{L}(k) \geq \mathcal{L}(k-1)$ for $k \geq \hat{N}$. 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.

13.18 (a) Setting the derivative of the likelihood 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

$$
\begin{aligned}
\mathcal{I}(N) &= -E_m\left[\frac{\partial^2}{\partial N^2}\log \mathcal{L}(N/m, n_1, n_2)\right] \\
&= -E_m\left[\frac{mN^2 - 2Nn_1n_2 + n_1^2 n_2}{N^2(N-n_1)^2}\right]
\end{aligned}
$$

$$\begin{aligned} &= -\frac{Nn_1n_2 - 2Nn_1n_2 + n_1^2n_2}{N^2(N-n_1)^2} \\ &= \frac{n_1n_2}{N^2(N-n_1)}. \end{aligned}$$

Consequently, the asymptotic variance of $\hat{N}$ is

$$V(\hat{N}) = \frac{1}{\mathcal{I}(N)} = \frac{N^2(N-n_1)}{n_1n_2}.$$

13.19 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

$$n_2 = C - N + \sqrt{N(N-C)}$$

13.20 (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 $q = n_2 + 1$

$$E[\tilde{N}] = E\left[\frac{(n_1+1)(n_2+1)}{X+1} - 1\right]$$

$$
\begin{aligned}
&= \sum_{m=0}^{n_2} \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_2}} \frac{(n_1+1)(n_2+1)}{m+1} - 1 \\
&= \sum_{m=0}^{n_2} (N+1) \frac{\binom{n_1+1}{m+1}\binom{N+1-(n_1+1)}{n_2+1-(m+1)}}{\binom{N+1}{n_2+1}} - 1 \\
&= (N+1) \sum_{k=1}^{q} \frac{\binom{n_1+1}{k}\binom{N-n_1}{q-k}}{\binom{N+1}{q}} - 1 \\
&= (N+1) \left[\sum_{k=0}^{q} \frac{\binom{n_1+1}{k}\binom{N-n_1}{q-k}}{\binom{N+1}{q}} - \frac{\binom{N-n_1}{q}}{\binom{N+1}{q}} \right] - 1.
\end{aligned}
$$

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 $n_2 \geq N - n_1$, $E[\tilde{N}] = N$.

13.21 Answers will vary.

Chapter 14

Rare Populations and Small Area Estimation

14.1 Answers will vary. All of these are rare populations in the sense discussed in the chapter, even if the population size may be large. For example, according to Canada's 2011 National Household Survey there were approximately one million Muslims in Canada in 2011 (about 3.2% of the population). But there is no list of them.

See `https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/Presentations/2015_FCSM/H3_Young_2015FCSM.pdf`
and
`https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/Presentations/2018_FCSM_Research_and_Policy_Conference/Evaluating_the_Use_of_Web-Scraped_List_Frames_to_Assess_Undercoverage_in_Surveys.pdf`
for more information about urban agriculture.

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 $A = \frac{N_1 p_1}{Np}$, $1 - A = \frac{N_2 p_2}{Np}$, $n_1 p_1 = k f_2 p_1 N_1$, $n_2 p_2 = f_2 p_2 N_2$, and

$$\begin{aligned} V(\hat{\bar{y}}_d) &\approx \frac{S_1^2}{f_2}\left(\frac{A^2}{kN_1p_1} + \frac{(1-A)^2}{N_2p_2}\right) \\ &= S_1^2\left[\left(\frac{N_1p_1}{Np}\right)^2 \frac{1}{kf_2N_1p_1} + \left(\frac{N_2p_2}{Np}\right)^2 \frac{1}{f_2N_1p_1}\right] \\ &= \frac{S_1^2}{(Np)^2 f_2}\left(\frac{N_1p_1}{k} + N_2p_2\right). \end{aligned}$$

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

$$\begin{aligned} V(\hat{\bar{y}}_d) &\approx \frac{S_1^2}{(Np)^2 f_2}\left(\frac{N_1p_1}{k} + N_2p_2\right) \\ &= \frac{S_1^2}{(Np)^2 n}(N_1k + N_2)\left(\frac{N_1p_1}{k} + N_2p_2\right), \end{aligned}$$

or, equivalently, to minimize

$$g(k) = (N_1k + N_2)\left(\frac{N_1p_1}{k} + N_2p_2\right).$$

Setting the derivative

$$\frac{dg}{dk} = N_1N_2p_2 - \frac{N_1N_2p_1}{k^2}$$

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 θ.

$$\begin{aligned} \frac{d}{d\theta}V(\hat{t}_{y,\theta}) &= \frac{d}{d\theta}\Big\{V\left[\hat{t}_a^A\right] + \theta^2 V\left[\hat{t}_{ab}^A\right] + 2\theta\text{Cov}\left[\hat{t}_a^A, \hat{t}_{ab}^A\right] \\ &\qquad +(1-\theta)^2 V\left[\hat{t}_{ab}^B\right] + V\left[\hat{t}_b^B\right] + 2(1-\theta)\text{Cov}\left[\hat{t}_b^B, \hat{t}_{ab}^B\right]\Big\} \\ &= 2\theta V\left[\hat{t}_{ab}^A\right] + 2\text{Cov}\left[\hat{t}_a^A, \hat{t}_{ab}^A\right] - 2(1-\theta)V\left[\hat{t}_{ab}^B\right] - 2\text{Cov}\left[\hat{t}_b^B, \hat{t}_{ab}^B\right] \end{aligned}$$

Setting the derivative equal to 0 and solving gives the optimal value of θ.

14.4 First, note that because samples are selected independently from frames A and B,

$$\begin{aligned} V\left[\hat{t}_a^A + \alpha\hat{t}_{ab}^A + (1-\alpha)\hat{t}_{ab}^B + \hat{t}_b^B\right] &= V\left[\hat{t}_a^A + \alpha\hat{t}_{ab}^A\right] + V\left[(1-\alpha)\hat{t}_{ab}^B + \hat{t}_b^B\right] \\ &= \frac{V_A}{n_A} + \frac{V_B}{n_B}. \end{aligned}$$

Now use Lagrange multipliers to minimize $\frac{V_A}{n_A}+\frac{V_B}{n_B}$ subject to the constraint that $C_0 = c_A n_A + c_B n_B$. Let

$$g(n_A, n_B, \lambda) = \frac{V_A}{n_A} + \frac{V_B}{n_B} + \lambda\left[c_A n_A + c_B n_B - C_0\right].$$

The partial derivatives are:

$$\frac{\partial g}{\partial n_A} = -\frac{V_A}{n_A^2} + \lambda c_A,$$

$$\frac{\partial g}{\partial n_B} = -\frac{V_B}{n_B^2} + \lambda c_B,$$

$$\frac{\partial g}{\partial \lambda} = c_A n_A + c_B n_B - C_0.$$

Setting the partial derivatives equal to zero, we have:

$$\frac{V_A}{n_A^2} = \lambda c_A, \tag{14.1}$$

$$\frac{V_B}{n_B^2} = \lambda c_B, \tag{14.2}$$

$$C_0 = c_A n_A + c_B n_B.$$

This implies that

$$c_A n_A = \sqrt{\frac{c_A V_A}{\lambda}}$$

and

$$c_B n_B = \sqrt{\frac{c_B V_B}{\lambda}},$$

which, since $c_A n_A + c_B n_B = C_0$, implies that

$$\sqrt{\lambda} = \frac{1}{C_0}\left(\sqrt{c_A V_A} + \sqrt{c_B V_B}\right).$$

Consequently, the variance is minimized when

$$n_A = C_0 \frac{\sqrt{V_A/c_A}}{\sqrt{c_A V_A} + \sqrt{c_B V_B}}$$

and

$$n_B = C_0 \frac{\sqrt{V_B/c_B}}{\sqrt{c_A V_A} + \sqrt{c_B V_B}}.$$

An alternative approach is to think of this as an allocation problem for stratified sampling, and use results from Chapter 3 to obtain the result without redoing the Lagrange multiplier argument.

14.8 (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

$$\begin{aligned} V[\tilde{\theta}_d(\alpha_d) - \theta_d] &= (\alpha_d - 1)^2\sigma_v^2 + \alpha_d^2\psi_d \\ &= \left(\frac{\psi_d}{\sigma_v^2 + \psi_d}\right)^2 \sigma_v^2 + \left(\frac{\sigma_v^2}{\sigma_v^2 + \psi_d}\right)^2 \psi_d \\ &= \frac{\psi_d^2\sigma_v^2}{(\sigma_v^2 + \psi_d)^2} + \frac{\psi_d\sigma_v^4}{(\sigma_v^2 + \psi_d)^2} \\ &= \frac{\psi_d\sigma_v^2(\psi_d + \sigma_v^2)}{(\sigma_v^2 + \psi_d)^2} \\ &= \alpha_d\psi_d. \end{aligned}$$

14.9 Here is SAS code for construction the population and samples:

```
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;
```

Here is code for R:

```
ex09<-data.frame(stratum=rep(1:20,each=480/20),
                 psu=rep(1:8,times=480/8))
ex09$y=ex09$stratum
ex09$domain <- rep(1,nrow(ex09))
ex09$domain[ex09$psu>=5] <- 2
# select sample of 2 psus per stratum
```

```
set.seed(32827)
index <- NULL
for (i in 1:20) {
   stratsmp<-sample(1:8,size=2,replace=F)
   index <- c(index,which(ex09$stratum==i &
             (ex09$psu==stratsmp[1] | ex09$psu==stratsmp[2])))
}
ex09samp<-ex09[index,]
# Now look at domain statistics
dex09<-svydesign(id=~psu,strata=~stratum,nest=TRUE,data=ex09samp)
dex09dom1<-subset(dex09,domain==1)
svymean(~y,dex09dom1)
```

14.10 Contact tracing is a form of snowball sampling. The biggest source of potential bias is that the starting list of persons with the disease is biased. For COVID-19, for example, some communities had very low levels of testing and persons in those communities had a very low chance of occurring in a snowball sample.

A better method to estimate the prevalence of a disease is to take a probability sample and test persons in that sample, as in Example 6.12. This gives approximately unbiased estimates of the prevalence and of the characteristics of persons with the disease. It is also, however, far more expensive and, if the disease is rare, devotes a lot of the survey budget to persons without the disease.

Chapter 15

Nonprobability Samples

15.1 Answers will vary, but here are my answers.

(a) This is a judgment sample. The investigators deliberately selected a sample to mimic the census with respect to these variables. But who knows whether the sample is like the population with respect to sleep difficulties? Perhaps people with sleep difficulties are more likely to sign up for a sleep study.

(b) Convenience sample. But the author took this sample for the purpose of teaching statistics to 4th-graders, and it is fine for that purpose. No one is generalizing to another population.

(c) This is a census. They looked at all of the papers. But it seems to be unnecessary effort—with a probability sample, they could have taken a smaller sample or looked at a wider range of journals.

(d) This is a purposive, or judgment sample. A similar method was used for monitoring the COVID-19 pandemic in the U.S. It provides information on the same localities over time and thus gives comparisons of one flu season to the next. But it is not guaranteed to be representative.

(e) Quota sample, where the quota classes are the different race/ethnicity groups. Women responding to the solicitations may well differ from other postmenopausal women—for example, they may be more likely to have symptoms.

(f) This was a real study (Mccallum and Bury, 2013). But internet searches are not representative of public opinion. You cannot conclude anything from a decreasing frequency of web searches.

(g) These are administrative records. They describe nearly everyone in England, and serve as a database for studying public health.

(h) This is a convenience sample, since people self-select whether to participate. Whether it provides a suitable sample depends on how similar the self-selected sample is to the population.

15.2 The *New York Times* uses a purposive sample to determine the Bestseller list. For the list to be an accurate representation of sales, the retailers need to be representative of the set of all book retailers. It is not clear that this is the case. And it is not representative of readership, because no one knows whether someone who buys a book reads it.

Also, the list can be manipulated. There have been cases in which an organization bought tens of thousands of books by an author in an effort to catapult the book onto the bestseller list; the newspaper puts a dagger next to those titles to indicate the large number of institutional purchases.

15.3 Answers will vary.

15.4 This is a convenience sample. No matter what type of opt-in process is used, the sample still consists entirely of volunteers. Biases could go in almost any direction, but one point is clear: the statistic from the survey describes the set of persons who took the sample, but can not be generalized to a population beyond that sample.

15.9 (a) The contingency table is given below, along with the percentage who are female in each category of *prof_cat*.

	Male Count	Female Count	Percent Female
Novice	320	360	52.94118
Average	546	739	57.50973
Professional	210	200	48.78049

A χ^2 test of association has $X^2 = 10.74$. Comparing this to a χ^2 distribution with 2 df yields p-value = 0.0047. Men are more likely to be professional respondents. If you collapse the novice and average categories and perform a chi-square test on the 2×2 table, $X^2 = 6.99$ and the p-value is 0.0082. Either way, men are significantly more likely to be professional respondents than women.

(b) The table below gives the age distribution for the three categories of *prof_cat*.

	18–34	35–49	50–64	65+
Novice	354	203	89	34
Average	347	412	329	197
Professional	49	96	163	102
Total	750	711	581	333
Percent professional	6.53	13.50	28.06	30.63

Older age groups have much higher percentages of professional respondents. To test association in

the 3×4 table, $X^2 = 307.97$; comparing to a χ^2_6 distribution gives p-value < 0.0001. If conducting a test with the novice and average groups combined into one category, then $X^2 = 156.5318$; comparing to a χ^2 distribution with 3 df gives p-value $< .0001$.

(c) The percentages of persons who are motivated by money are:

Novice 37.997%; Average 52.529 %; Professional 67.073%.

$X^2 = 89.507$; comparing to a χ^2 distribution with 2 df, the p-value is way less than 0.0001.

(d) Men are more likely to be professional respondents and professional respondents are more likely to be motivated by money. But, interestingly, men are less likely to be motivated by money (it's a nice instance of Simpson's paradox).

Also, older age groups are more likely to be professional respondents.

Any volunteer online panel has a potential for bias, because persons self-select to participate. For this one, apparently, many of them are motivated to participate for money. This does not necessarily mean that they will give biased or misleading responses—indeed, Zhang et al. (2020) found that the "professional" respondents gave higher quality data than many of the novice respondents. But being a professional respondent was associated with age and gender, and this could lead to bias for some estimates.

15.10 (a) Here is the table of unweighted percentages from the survey.

	Survey Percentage	ACS Percentage
Women	54.58	51.43
College Graduates	36.77	25.98
Age 18 to 34	31.82	30.61
Age 35 to 64	54.20	51.98
Age 65 and over	13.98	17.41

The survey respondents are disproportionately female and more highly educated. Senior citizens are underrepresented among the survey respondents.

(b) The following code can be used to create poststratified weights for the survey and calculate the estimates. The variable *wt* has value 1 for all observations.

```
proc surveymeans data=profresp mean clm;
   class panelq1 motive non_white prof_cat;
   weight wt;
   poststrata gender age3cat edu3cat/ PSTOTAL = profrespACS OUTPSWGT = pswgt;
   var panelq1 motive non_white prof_cat motive freq_q1-freq_q5;
run;
```

The coefficient of variation of the poststratified weights is 0.632.

(c) Here are the unweighted estimates:

Variable	Level	Mean	Std Error of Mean	95% CL for Mean	
freq_q1		3.670706	0.130003	3.41577669	3.92563577
freq_q2		8.122748	0.188266	7.75356689	8.49192955
freq_q3		5.715539	0.132116	5.45646466	5.97461303
freq_q4		6.397746	0.104824	6.19219145	6.60330103
freq_q5		5.455833	0.370045	4.73019229	6.18147437
panelq1	1	0.646199	0.009774	0.62703161	0.66536605
	2	0.353801	0.009774	0.33463395	0.37296839
motive	1	0.269231	0.009071	0.25144262	0.28701891
	2	0.168896	0.007662	0.15387127	0.18392137
	3	0.508361	0.010224	0.48831245	0.52840996
	4	0.053512	0.004602	0.04448644	0.06253698
non_white	0	0.815712	0.007927	0.80016703	0.83125796
	1	0.184288	0.007927	0.16874204	0.19983297
prof_cat	1	0.286316	0.009278	0.26812277	0.30450881
	2	0.541053	0.010227	0.52099727	0.56110799
	3	0.172632	0.007757	0.15742124	0.18784191

Here are the weighted estimates:

Variable	Level	Mean	Std Error of Mean	95% CL for Mean	
freq_q1		3.721616	0.141574	3.44399507	3.99923740
freq_q2		8.028034	0.221072	7.59452103	8.46154670
freq_q3		5.612264	0.185379	5.24874417	5.97578305
freq_q4		6.264698	0.121001	6.02741997	6.50197657
freq_q5		5.699053	0.723754	4.27980516	7.11830164
panelq1	1	0.662629	0.010469	0.64209962	0.68315903
	2	0.337371	0.010469	0.31684097	0.35790038
motive	1	0.266301	0.010592	0.24553083	0.28707085
	2	0.170706	0.009157	0.15274878	0.18866317
	3	0.502518	0.012009	0.47896918	0.52606622
	4	0.060475	0.006139	0.04843646	0.07251450
non_white	0	0.820579	0.008727	0.80346513	0.83769340
	1	0.179421	0.008727	0.16230660	0.19653487
prof_cat	1	0.274682	0.009727	0.25560752	0.29375579

```
2 0.542434 0.011975 0.51895057 0.56591738
3 0.182884 0.009443 0.16436662 0.20140212
```

For this example, the two sets of estimates are similar.

(d) But this does not mean that the estimates of questions 1 to 5 give accurate estimates of the corresponding population quantities. The only way of assessing the accuracy of the estimates is to obtain information from an external source. For example, one might check NHANES for 2011 to look at the mean number of doctor visits per year from that survey.

15.11 Using the spreadsheet in `hunting.csv`, we calculate $\sum_{i\in\mathcal{S}} w_i^2$ =1995 and $\sum_{i\in\mathcal{S}} w_i = 1212$. Thus, using Equation (7.12), the variance inflation due to weights is 1995/1212 = 1.645. The margin of error is 0.036.

15.12 The statistics in this example were from an antibody test for COVID-19. Results for this exercise are discussed in `https://www.sharonlohr.com/blog/2020/5/11/taking-an-antibody-test`, which works through the steps in a table.

(a) The weight for each unit in Sample A should be 5000/108 = 46.2963. The weight for each unit in Sample B should be 95000/433 = 219.3995.

(b) The table entries, with 95% CIs in square brackets behind the entries, are given in the following table. The entries are calculated as follows: $93303 = 428 \times 219.3995$; $1097 = 5 \times 219.3995$; $602 = 13 \times 46.2963$; $5495 = 95 \times 46.2963$.

	Test Negative for X	Test Positive for X	
Don't Have X	93,903 [91,543, 95,767]	1,097 [357, 2,541]	95,000
Have X	602 [136, 1,682]	4,398 [2,832, 6,484]	5,000
	94,505 [92,236, 96,269]	5,495	100,000

(c) The positive predictive value is easily calculated from the table in part (b) as 4398/5495, or 80%. The Clopper-Pearson 95% CI is [61.7%, 92.2%].

(d) This method assumes that the persons in the two samples are essentially simple random samples of the people with, and without, the antibodies in the population of interest. That is a strong assumption, since these are convenience samples.

It is desirable in practice to have multiple laboratories assess the sensitivity and specificity of a proposed diagnostic test: different labs will conduct the assessments on samples from different sets of people, which gives one more confidence in the results.

Here is code in SAS software for performing the calculations:

```
%let popsize = 100000; /* Can use any population size for calculating probabilities */
%let prevalence = 5; /* Estimate positive predictive value under assumed prevalence 5% */

data assay_count;
   input true_status $ test_status $ count;
   if true_status = 'No_X' then wgt = &popsize * (100 - &prevalence) / (100 * 433) ;
   else wgt = &popsize * &prevalence / (100 * 108);
   datalines;
   No_X Test_Neg 428
   No_X Test_Pos 5
   X    Test_Neg 13
   X    Test_Pos 95
;
run;

data assay;
   set assay_count;
   do i = 1 to count;
      output;
   end;
run;

proc surveyfreq data=assay;
        tables true_status*test_status / cl (type = cp) row column;
        weight wgt;
run;
```

15.13 Let's first find the population variance of x and y. In the notation of the contingency table, $\bar{x}_{\mathcal{U}} = N_{1+}/N$, $\sum_{i=1}^{N} x_i = N_{1+}$, $\bar{y}_{\mathcal{U}} = N_{+1}/N$, and $\sum_{i=1}^{N} y_i = N_{+1}$.

$$
\begin{aligned}
S_x^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x}_{\mathcal{U}})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[x_i^2 - 2x_i\bar{x}_{\mathcal{U}} + \bar{x}_{\mathcal{U}}^2\right] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[x_i - 2x_i\frac{N_{1+}}{N} + \left(\frac{N_{1+}}{N}\right)^2\right] \\
&= \frac{1}{N-1}\left(N_{1+} - 2\frac{N_{1+}^2}{N} + \frac{N_{1+}^2}{N}\right) \\
&= \frac{N_{1+}}{N-1}\left(1 - \frac{N_{1+}}{N}\right). \qquad (15.1)
\end{aligned}
$$

Similarly, $S_y^2 = \frac{N_{+1}}{N-1}\left(1 - \frac{N_{+1}}{N}\right)$. Also,

$$\sum_{i=1}^{N}(x_i - \bar{x}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}_{\mathcal{U}}\bar{y}_{\mathcal{U}} = N_{11} - \frac{N_{1+}N_{+1}}{N}.$$

Thus,

$$\begin{aligned}
R &= \frac{\sum_{i=1}^{N}(x_i - \bar{x}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})}{(N-1)S_x S_y} \\
&= \frac{N_{11} - \frac{N_{1+}N_{+1}}{N}}{(N-1)\sqrt{\frac{N_{1+}}{N-1}\left(1 - \frac{N_{1+}}{N}\right)\frac{N_{+1}}{N-1}\left(1 - \frac{N_{+1}}{N}\right)}} \\
&= \frac{NN_{11} - N_{1+}N_{+1}}{\sqrt{N_{1+}(N - N_{1+})N_{+1}(N - N_{+1})}} \\
&= \frac{(N_{00} + N_{01} + N_{10} + N_{11})N_{11} - (N_{10} + N_{11})(N_{01} + N_{11})}{\sqrt{N_{1+}(N - N_{1+})N_{+1}(N - N_{+1})}} \\
&= \frac{N_{00}N_{11} + N_{01}N_{11} + N_{10}N_{11} + N_{11}^2 - N_{10}N_{01} - N_{11}N_{01} - N_{10}N_{11} - N_{11}^2}{\sqrt{N_{1+}N_{0+}N_{+1}N_{+0}}} \\
&= \frac{N_{00}N_{11} - N_{10}N_{01}}{\sqrt{N_{1+}N_{0+}N_{+1}N_{+0}}}.
\end{aligned}$$

Note that R is related to Pearson's chi-squared statistic for testing the independence of x and y. From Equation (10.2) of SDA (applied to the population),

$$\begin{aligned}
X^2 &= \frac{\left(N_{00} - \frac{N_{0+}N_{+0}}{N}\right)^2}{\frac{N_{0+}N_{+0}}{N}} + \frac{\left(N_{01} - \frac{N_{0+}N_{+1}}{N}\right)^2}{\frac{N_{0+}N_{+1}}{N}} + \frac{\left(N_{10} - \frac{N_{1+}N_{+0}}{N}\right)^2}{\frac{N_{1+}N_{+0}}{N}} + \frac{\left(N_{11} - \frac{N_{1+}N_{+1}}{N}\right)^2}{\frac{N_{1+}N_{+1}}{N}} \\
&= \frac{(NN_{00} - N_{0+}N_{+0})^2}{NN_{0+}N_{+0}} + \frac{(NN_{01} - N_{0+}N_{+1})^2}{NN_{0+}N_{+1}} + \frac{(NN_{10} - N_{1+}N_{+0})^2}{NN_{1+}N_{+0}} + \frac{(NN_{11} - N_{1+}N_{+1})^2}{NN_{1+}N_{+1}} \\
&= \frac{N_{1+}N_{+1}(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}} + \frac{N_{1+}N_{+0}(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}} \\
&\quad + \frac{N_{0+}N_{+1}(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}} + \frac{N_{0+}N_{+0}(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}} \\
&= \frac{(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}}(N_{1+}N_{+1} + N_{1+}N_{+0} + N_{0+}N_{+1} + N_{0+}N_{+0}) \\
&= \frac{(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}}[N_{1+}N_{+1} + N_{1+}(N - N_{+1}) + (N - N_{1+})N_{+1} + (N - N_{1+})(N - N_{+1})] \\
&= \frac{(N_{00}N_{11} - N_{01}N_{10})^2}{NN_{0+}N_{+0}N_{1+}N_{+1}}[N_{1+}N_{+1} + N_{1+}(N - N_{+1}) + (N - N_{1+})N_{+1} + (N - N_{1+})(N - N_{+1})]
\end{aligned}$$

$$= \frac{N\left(N_{00}N_{11} - N_{01}N_{10}\right)^2}{N_{0+}N_{+0}N_{1+}N_{+1}}$$
$$= NR^2.$$

15.14. Note that if $S_y^2 = 0$, then $y_i = \bar{y}_{\mathcal{U}}$ for all units i and $\bar{y}_{\mathcal{C}} - \bar{y}_{\mathcal{U}} = 0$. Similarly, if $n = N$, then $\bar{y}_{\mathcal{C}} = \bar{y}_{\mathcal{U}}$ and the error is zero.

In the following, we assume $n < N$ and $S_y^2 > 0$ so there is no division by zero.

$$\begin{aligned}
\bar{y}_{\mathcal{C}} - \bar{y}_{\mathcal{U}} &= \frac{\sum_{i=1}^N R_i y_i}{\sum_{i=1}^N R_i} - \bar{y}_{\mathcal{U}} \\
&= \frac{\sum_{i=1}^N R_i y_i - \left(\sum_{i=1}^N R_i\right)\bar{y}_{\mathcal{U}}}{\sum_{i=1}^N R_i} \\
&= \frac{\sum_{i=1}^N R_i \left(y_i - \bar{y}_{\mathcal{U}}\right)}{\sum_{i=1}^N R_i} \\
&= \frac{\sum_{i=1}^N \left(R_i - \bar{R}_{\mathcal{U}}\right)\left(y_i - \bar{y}_{\mathcal{U}}\right)}{n}.
\end{aligned} \tag{15.2}$$

The last line is true because $\sum_{i=1}^N \bar{R}_{\mathcal{U}}(y_i - \bar{y}_{\mathcal{U}}) = \bar{R}_{\mathcal{U}} \sum_{i=1}^N (y_i - \bar{y}_{\mathcal{U}}) = 0$ and because $\sum_{i=1}^N R_i = n$.

We can now use the definition of the population correlation coefficient in Equation (A.10) of SDA to rewrite the error in terms of population correlations and variances. First, calculate S_R^2, noting that $\bar{R}_{\mathcal{U}} = n/N$:

$$\begin{aligned}
S_R^2 &= \frac{1}{N-1}\sum_{i=1}^N (R_i - \bar{R}_{\mathcal{U}})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^N \left[R_i^2 - 2R_i\bar{R}_{\mathcal{U}} + \bar{R}_{\mathcal{U}}^2\right] \\
&= \frac{1}{N-1}\sum_{i=1}^N \left[R_i - 2R_i\frac{n}{N} + \left(\frac{n}{N}\right)^2\right] \\
&= \frac{1}{N-1}\left(n - 2\frac{n^2}{N} + \frac{n^2}{N}\right) \\
&= \frac{n}{N-1}\left(1 - \frac{n}{N}\right).
\end{aligned} \tag{15.3}$$

Then,

$$\bar{y}_{\mathcal{C}} - \bar{y}_{\mathcal{U}} = \frac{\sum_{i=1}^N \left(R_i - \bar{R}_{\mathcal{U}}\right)\left(y_i - \bar{y}_{\mathcal{U}}\right)}{n}$$

$$= \text{Corr}\,(R, y) S_R S_y \frac{N-1}{n}$$
$$= \text{Corr}\,(R, y) \sqrt{\frac{N-1}{n}\left(1 - \frac{n}{N}\right)} S_y.$$

15.15 Because the sample size is fixed, $\sum_{i=1}^{N} Z_i = n$ and $\bar{Z}_{\mathcal{U}} = n/N$. We also use the expression for S_R^2 in (SM15.3) that

$$S_R^2 = S_Z^2 = \frac{n}{N-1}\left(1 - \frac{n}{N}\right).$$

Thus, S_Z^2 contains no random variables and can be moved outside of expectations.

(a) The expected value of the correlation is

$$E\left[\text{Corr}\,(Z, y)\right] = E\left[\frac{\sum_{i=1}^{N}(Z_i - \bar{Z}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})}{(N-1)S_y S_Z}\right]$$
$$= \frac{1}{(N-1)S_y S_Z}\sum_{i=1}^{N}(y_i - \bar{y}_{\mathcal{U}}) E\left[Z_i - \bar{Z}_{\mathcal{U}}\right]$$
$$= 0.$$

(b) The easiest way to prove this result is to use the already proven result about the variance (= mean squared error, because the sample mean from an SRS is unbiased) of the sample mean from an SRS from Chapter 2: $V(\bar{y}) = (S_y^2/n)(1 - n/N)$. We equate the two mean squared error expressions to get:

$$\frac{S_y^2}{n}\left(1 - \frac{n}{N}\right) = E\left[\text{Corr}^{\,2}(Z, y)\right]\frac{N-1}{n}\left(1 - \frac{n}{N}\right) S_y^2.$$

Solving for $E\left[\text{Corr}^{\,2}(R, y)\right]$ gives

$$E\left[\text{Corr}^{\,2}(Z, y)\right] = \frac{1}{N-1}.$$

Or, one can work through the expected values directly using the random variables. The expected value of the square of the correlation is

$$E\left[\text{Corr}^{\,2}(Z, y)\right] = E\left[\frac{\sum_{i=1}^{N}(Z_i - \bar{Z}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})\sum_{j=1}^{N}(Z_j - \bar{Z}_{\mathcal{U}})(y_j - \bar{y}_{\mathcal{U}})}{(N-1)^2 S_y^2 S_Z^2}\right]$$
$$= \frac{1}{n(N-1)S_y^2\left(1 - \frac{n}{N}\right)}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - \bar{y}_{\mathcal{U}})(y_j - \bar{y}_{\mathcal{U}}) E\left[Z_i Z_j - Z_i\frac{n}{N} - Z_j\frac{n}{N} + \frac{n^2}{N^2}\right]$$
$$= \frac{1}{n(N-1)S_y^2\left(1 - \frac{n}{N}\right)}\left\{\sum_{i=1}^{N}(y_i - \bar{y}_{\mathcal{U}})^2 E\left[Z_i^2 - 2Z_i\frac{n}{N} + \frac{n^2}{N^2}\right]\right.$$

$$
\begin{aligned}
&\quad + \sum_{i=1}^{N}\sum_{j\neq i}^{N}(y_i-\bar{y}_{\mathcal{U}})(y_j-\bar{y}_{\mathcal{U}})E\left[Z_iZ_j - Z_i\frac{n}{N} - Z_j\frac{n}{N} + \frac{n^2}{N^2}\right]\Bigg\} \\
&= \frac{1}{n(N-1)S_y^2\left(1-\frac{n}{N}\right)}\left\{\sum_{i=1}^{N}(y_i-\bar{y}_{\mathcal{U}})^2\left[\frac{n}{N}-\frac{n^2}{N^2}\right]\right. \\
&\quad \left. + \sum_{i=1}^{N}\sum_{j\neq i}^{N}(y_i-\bar{y}_{\mathcal{U}})(y_j-\bar{y}_{\mathcal{U}})\left[\frac{n(n-1)}{N(N-1)}-\frac{n^2}{N^2}\right]\right\} \\
&= \frac{1}{n(N-1)S_y^2\left(1-\frac{n}{N}\right)}\left\{\sum_{i=1}^{N}(y_i-\bar{y}_{\mathcal{U}})^2\left[\frac{n}{N}-\frac{n^2}{N^2}\right]\right. \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N}(y_i-\bar{y}_{\mathcal{U}})(y_j-\bar{y}_{\mathcal{U}})\left[\frac{n(n-1)}{N(N-1)}-\frac{n^2}{N^2}\right] \\
&\quad \left. - \sum_{i=1}^{N}(y_i-\bar{y}_{\mathcal{U}})^2\left[\frac{n(n-1)}{N(N-1)}-\frac{n^2}{N^2}\right]\right\} \\
&= \frac{1}{n(N-1)S_y^2\left(1-\frac{n}{N}\right)}\left\{\sum_{i=1}^{N}(y_i-\bar{y}_{\mathcal{U}})^2\left[\frac{n}{N}-\frac{n^2}{N^2}\right]\right. \\
&\quad \left. - \sum_{i=1}^{N}(y_i-\bar{y}_{\mathcal{U}})^2\left[\frac{n(n-1)}{N(N-1)}-\frac{n^2}{N^2}\right]\right\} \\
&= \frac{1}{n(N-1)S_y^2\left(1-\frac{n}{N}\right)}\left\{\sum_{i=1}^{N}(y_i-\bar{y}_{\mathcal{U}})^2\right\}\left[\frac{n}{N}-\frac{n^2}{N^2}-\frac{n(n-1)}{N(N-1)}+\frac{n^2}{N^2}\right] \\
&= \frac{1}{n(N-1)S_y^2\left(1-\frac{n}{N}\right)}(N-1)S_y^2\frac{n}{N}\left(1-\frac{n-1}{N-1}\right) \\
&= \frac{1}{N-1}.
\end{aligned}
$$

15.16. *Coverage probability of a confidence interval when there is bias.*

(a) Standardize the random variable so it has a standard normal distribution. In the following, $T \sim N(0,1)$.

$$
\begin{aligned}
P\left(|\bar{y}_{\mathcal{C}}-\bar{y}_{\mathcal{U}}| \leq 1.96\sqrt{v}\right) &= P\left(\frac{|\bar{y}_{\mathcal{C}}-\bar{y}_{\mathcal{U}}|}{\sqrt{v}} \leq 1.96\right) \\
&= P\left(\frac{|\bar{y}_{\mathcal{C}}-E[\bar{y}_{\mathcal{C}}]+b|}{\sqrt{v}} \leq 1.96\right) \\
&= 1-P\left(\frac{|\bar{y}_{\mathcal{C}}-E[\bar{y}_{\mathcal{C}}]+b|}{\sqrt{v}} > 1.96\right) \\
&= 1-P\left(\frac{\bar{y}_{\mathcal{C}}-E[\bar{y}_{\mathcal{C}}]+b}{\sqrt{v}} > 1.96\right) - P\left(\frac{-\left(\bar{y}_{\mathcal{C}}-E[\bar{y}_{\mathcal{C}}]+b\right)}{\sqrt{v}} > 1.96\right)
\end{aligned}
$$

$$
\begin{aligned}
&= 1 - P\left(\frac{\bar{y}_{\mathcal{C}} - E[\bar{y}_{\mathcal{C}}]}{\sqrt{v}} > -\frac{b}{\sqrt{v}} + 1.96\right) - P\left(\frac{\bar{y}_{\mathcal{C}} - E[\bar{y}_{\mathcal{C}}]}{\sqrt{v}} < -\frac{b}{\sqrt{v}} - 1.96\right) \\
&= 1 - P\left(T > -\frac{b}{\sqrt{v}} + 1.96\right) - P\left(T < -\frac{b}{\sqrt{v}} - 1.96\right) \\
&= 1 - \frac{1}{\sqrt{2\pi}} \int_{-b/\sqrt{v}+1.96}^{\infty} e^{-t^2/2} dt - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-b/\sqrt{v}-1.96} e^{-t^2/2} dt.
\end{aligned}
$$

(b) Figure SM15.1 shows the plot of coverage percentage (100 × coverage probability) as a function of the relative bias $b/\sqrt{v}$. Note that for relatively small amounts of bias the coverage is close to 95%. As the bias increases relative to the variance, however, the coverage percentage becomes low indeed.

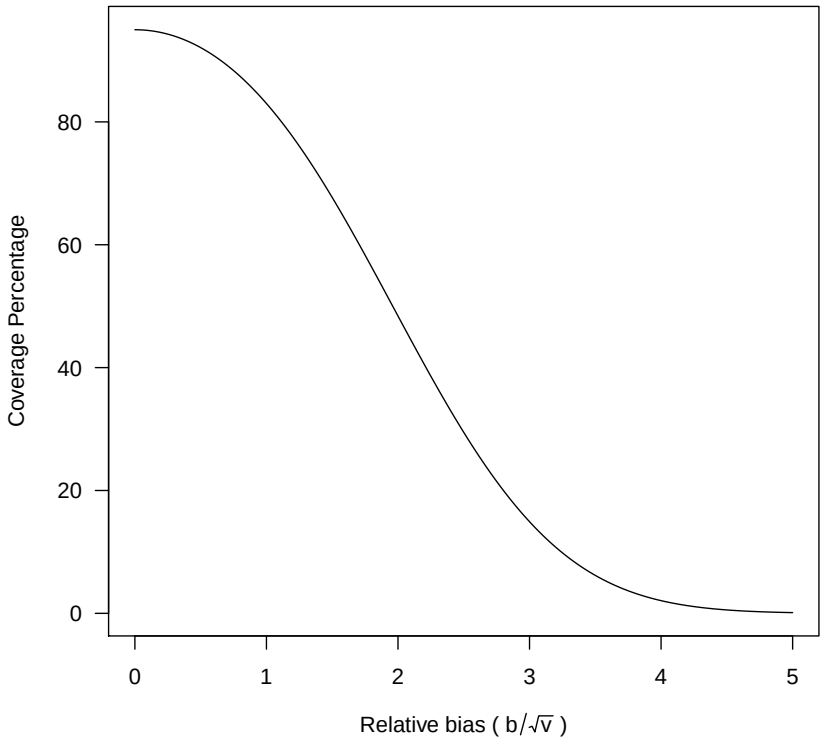


Figure 15.1: Plot of coverage percentage vs. relative bias.

(c) In ratio estimation, the bias is of order $1/n$ and decreases to zero as the sample size increases. For a nonprobability sample, however, the bias can remain roughly constant as the sample size increases. Thus, for large samples with bias, the bias term dominates the MSE and the confidence

interval will have incorrect coverage probability.

15.17 From (15.5),

$$\text{MSE}\left(\bar{y}_{\mathcal{C}}\right) = E_R\left[\text{Corr}^{\,2}(R, y)\right] \times \frac{N-1}{n}\left(1 - \frac{n}{N}\right) \times S_y^2$$

and, because the mean from an SRS is unbiased, from (2.12),

$$\text{MSE}\left(\bar{y}\right) = \frac{S_y^2}{n}\left(1 - \frac{n}{N}\right).$$

Thus, the ratio is

$$\frac{\text{MSE}\left(\bar{y}_{\mathcal{C}}\right)}{\text{MSE}\left(\bar{y}\right)} = (N-1)E_R\left[\text{Corr}^{\,2}(R, y)\right].$$

Even a small correlation between the response indicator and y is multiplied by $(N-1)$, so the design effect for a nonprobability sample can be immense. Note that the sample size n does not appear in the expression for the design effect—taking a larger sample does not help.

15.18 (a) The total sample size is n, so that must equal the sum of the number of responses obtained from survey participants. Each unit in the population must then participate somewhere between 0 and n times. If $Q_i = n$ for some i, then the sample mean is determined entirely by the response of one unit in the population.

(b) We have

$$\bar{y}_{\mathcal{Q}} - \bar{y}_{\mathcal{U}} = \frac{\sum_{i=1}^{N}\left(Q_i - \frac{n}{N}\right)\left(y_i - \bar{y}_{\mathcal{U}}\right)}{n} = \text{Corr}\left(Q, y\right)(N-1)S_Q S_y.$$

The variance S_Q^2 is:

$$S_Q^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Q_i - \bar{Q}_{\mathcal{U}}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Q_i - \frac{n}{N}\right)^2.$$

(c) Then an argument similar to that in (SM15.2), substituting Q_i for R_i, establishes the result. Because each value of Q_i is in the set $\{0, 1, \ldots, n\}$, Q_i^2 is greater than or equal to Q_i and

$$\sum_{i=1}^{N} Q_i^2 \geq \sum_{i=1}^{N} Q_i = n.$$

Thus,

$$(N-1)S_Q^2 = \sum_{i=1}^{N}\left(Q_i - \frac{n}{N}\right)^2$$

$$
\begin{aligned}
&= (N-1)\left\{\sum_{i=1}^{N} Q_i^2 - \frac{n^2}{N}\right\} \\
&\geq (N-1)\left\{n - \frac{n^2}{N}\right\} \\
&= (N-1)\left\{\sum_{i=1}^{N} R_i^2 - \frac{n^2}{N}\right\} \\
&= (N-1)S_R^2.
\end{aligned}
$$

Since $\text{Corr}\,(R, y) = \text{Corr}\,(Q, y)$ and the population standard deviation S_y is the same for both samples, this implies that

$$
\begin{aligned}
|\bar{y}_{\mathcal{Q}} - \bar{y}_{\mathcal{U}}| &= |\text{Corr}\,(Q, y)|\,(N-1)S_Q S_y \frac{1}{n} \\
&\geq |\text{Corr}\,(R, y)|\,(N-1)S_R S_y \frac{1}{n} \\
&= |\bar{y}_{\mathcal{C}} - \bar{y}_{\mathcal{U}}|\,.
\end{aligned}
$$

15.19 *Using a convenience sample to estimate population size.* First note that

$$
S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_{\mathcal{U}})^2 = \frac{N}{N-1}\bar{y}_{\mathcal{U}}(1-\bar{y}_{\mathcal{U}}),\; S_R^2 = \frac{N}{N-1}\bar{R}_{\mathcal{U}}(1-\bar{R}_{\mathcal{U}}).
$$

(a) We can use the arguments in Exercise 15.14 to obtain the desired result:

$$
\begin{aligned}
\sum_{i=1}^{N} R_i y_i - t_y &= \left(\sum_{i=1}^{N} R_i\right)\bar{y}_{\mathcal{C}} - N\bar{y}_{\mathcal{U}} \\
&= \left(\sum_{i=1}^{N} R_i\right)\{\bar{y}_{\mathcal{C}} - \bar{y}_{\mathcal{U}}\} - \left(N - \sum_{i=1}^{N} R_i\right)\bar{y}_{\mathcal{U}} \\
&= \left(\sum_{i=1}^{N} R_i\right)\{\bar{y}_{\mathcal{C}} - \bar{y}_{\mathcal{U}}\} - \left(N - \sum_{i=1}^{N} R_i\right)\bar{y}_{\mathcal{U}} \\
&= \sum_{i=1}^{N}\left(R_i - \bar{R}_{\mathcal{U}}\right)(y_i - \bar{y}_{\mathcal{U}}) - \left(N - \sum_{i=1}^{N} R_i\right)\bar{y}_{\mathcal{U}} \\
&= (N-1)\text{Corr}\,(R, y)S_R S_y - N\left(1 - \bar{R}_{\mathcal{U}}\right)\bar{y}_{\mathcal{U}} \\
&= N\text{Corr}\,(R, y)\sqrt{\bar{R}_{\mathcal{U}}(1-\bar{R}_{\mathcal{U}})\bar{y}_{\mathcal{U}}(1-\bar{y}_{\mathcal{U}})} - N\left(1 - \bar{R}_{\mathcal{U}}\right)\bar{y}_{\mathcal{U}}.
\end{aligned}
$$

(b) If we observe the entire population, then $\sum_{i=1} R_i y_i = \sum_{i=1}^{N} y_i$. Similarly, if no population members have the characteristic, then any sample catches all of them. For the third condition, set

the expression in (a) equal to zero and solve to obtain that the error is zero when

$$\text{Corr}\,(R, y) = \sqrt{\frac{\left(1 - \bar{R}_{\mathcal{U}}\right)\bar{y}_{\mathcal{U}}}{\left(1 - \bar{y}_{\mathcal{U}}\right)\bar{R}_{\mathcal{U}}}}. \tag{15.4}$$

The correlation attains the value on the right-hand-side of (SM15.4) if $\text{Corr}\,(R, y) = 1$ or if $R_i = 1$ whenever $y_i = 1$. In the latter case,

$$\begin{aligned}\sum_{i=1}^{N}(y_i - \bar{y}_{\mathcal{U}})(R_i - \bar{R}_{\mathcal{U}}) &= \sum_{i=1}^{N} y_i R_i - N\bar{y}_{\mathcal{U}}\bar{R}_{\mathcal{U}} \\ &= N\bar{y}_{\mathcal{U}} - N\bar{y}_{\mathcal{U}}\bar{R}_{\mathcal{U}} \\ &= N\bar{y}_{\mathcal{U}}(1 - \bar{R}_{\mathcal{U}}).\end{aligned}$$

15.20 *Prove Equation (15.7).* We follow the same arguments as in (SM15.2), substituting $T_i = R_i w_i$ for R_i. In the following, we assume $n < N$ and $S_y^2 > 0$ so there is no division by zero.

$$\begin{aligned}\bar{y}_{\mathcal{C}w} - \bar{y}_{\mathcal{U}} &= \frac{\sum_{i=1}^{N} R_i w_i y_i}{\sum_{i=1}^{N} R_i w_i} - \bar{y}_{\mathcal{U}} \\ &= \frac{\sum_{i=1}^{N} R_i w_i y_i - \left(\sum_{i=1}^{N} R_i w_i\right)\bar{y}_{\mathcal{U}}}{\sum_{i=1}^{N} R_i w_i} \\ &= \frac{\sum_{i=1}^{N} T_i\left(y_i - \bar{y}_{\mathcal{U}}\right)}{\sum_{i=1}^{N} R_i w_i} \\ &= \frac{\sum_{i=1}^{N}\left(T_i - \bar{T}_{\mathcal{U}}\right)\left(y_i - \bar{y}_{\mathcal{U}}\right)}{\sum_{i=1}^{N} R_i w_i} \\ &= \text{Corr}\,(T, y)(N-1)\frac{S_T}{\sum_{i=1}^{N} R_i w_i} S_y.\end{aligned} \tag{15.5}$$

Also, noting that $R_i^2 = R_i$ and that $(\sum_{i=1}^{N} R_i w_i)/n = \bar{w}_{\mathcal{C}}$, the average of the weights in the sample,

$$\begin{aligned}S_T^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(T_i - \bar{T}_{\mathcal{U}})^2 \\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{N}(R_i w_i^2) - \frac{\left(\sum_{i=1}^{N} R_i w_i\right)^2}{N}\right\} \\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{N} R_i\left(w_i^2 - \frac{\left(\sum_{j=1}^{N} R_j w_j\right)^2}{n^2} + \frac{\left(\sum_{j=1}^{N} R_j w_j\right)^2}{n^2}\right) - \frac{\left(\sum_{i=1}^{N} R_i w_i\right)^2}{N}\right\} \\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{N} R_i\left(w_i^2 - \frac{\left(\sum_{j=1}^{N} R_j w_j\right)^2}{n^2}\right) + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\sum_{i=1}^{N} R_i w_i\right)^2\right\}\end{aligned}$$

$$= \frac{1}{N-1}\left\{\sum_{i=1}^{N} R_i\left(w_i^2 - \bar{w}_{\mathcal{C}}^2\right) + \left(\frac{1}{n} - \frac{1}{N}\right) n^2 \bar{w}_{\mathcal{C}}^2\right\}.$$

Thus,

$$\begin{aligned}
\frac{S_T^2}{\left(\sum_{i=1}^{N} R_i w_i\right)^2} &= \frac{S_T^2}{n^2 \bar{w}_{\mathcal{C}}^2} \\
&= \frac{1}{N-1}\left\{\frac{\sum_{i=1}^{N} R_i\left(w_i^2 - \bar{w}_{\mathcal{C}}^2\right)}{n^2 \bar{w}_{\mathcal{C}}^2} + \left(\frac{1}{n} - \frac{1}{N}\right)\right\} \\
&= \frac{1}{n(N-1)}\left\{\frac{\sum_{i=1}^{N} R_i\left(w_i^2 - \bar{w}_{\mathcal{C}}^2\right)}{n \bar{w}_{\mathcal{C}}^2} + \left(1 - \frac{n}{N}\right)\right\} \\
&= \frac{1}{n(N-1)}\left\{\frac{n-1}{n}\mathrm{CV}_w^2 + \left(1 - \frac{n}{N}\right)\right\}
\end{aligned}$$

where

$$\mathrm{CV}_w = \frac{\sqrt{\sum_{i=1}^{N} R_i\left(w_i^2 - \bar{w}_{\mathcal{C}}^2\right)/(n-1)}}{\bar{w}_{\mathcal{C}}}$$

is the coefficient of variation of the weights in the sample.

Combining with (SM15.5), we have

$$\begin{aligned}
\bar{y}_{\mathcal{C}w} - \bar{y}_{\mathcal{U}} &= \mathrm{Corr}\,(T, y)(N-1)\frac{S_T}{\sum_{i=1}^{N} R_i w_i} S_y \\
&= \mathrm{Corr}\,(T, y)\sqrt{\frac{N-1}{n}\left\{\frac{n-1}{n}\mathrm{CV}_w^2 + \left(1 - \frac{n}{N}\right)\right\}} S_y.
\end{aligned}$$

15.21 Because the regression coefficients are the same, we have

$$\sum_{i \notin \mathcal{C}} y_i = \frac{\sum_{i \in \mathcal{C}} y_i}{\sum_{i \in \mathcal{C}} x_i} \sum_{i \notin \mathcal{C}} x_i.$$

Thus,

$$\begin{aligned}
\sum_{i \in \mathcal{C}} w_i y_i - t_y &= \frac{t_x}{\sum_{i \in \mathcal{C}} x_i} \sum_{i \in \mathcal{C}} y_i - \sum_{i=1}^{N} y_i \\
&= \frac{t_x}{\sum_{i \in \mathcal{C}} x_i} \sum_{i \in \mathcal{C}} y_i - \sum_{i \in \mathcal{C}} y_i - \sum_{i \notin \mathcal{C}} y_i \\
&= \frac{\sum_{i \in \mathcal{C}} x_i + \sum_{i \notin \mathcal{C}} x_i}{\sum_{i \in \mathcal{C}} x_i} \sum_{i \in \mathcal{C}} y_i - \sum_{i \in \mathcal{C}} y_i - \frac{\sum_{i \in \mathcal{C}} y_i}{\sum_{i \in \mathcal{C}} x_i} \sum_{i \notin \mathcal{C}} x_i
\end{aligned}$$

$$= 0.$$

15.22 Suppose that $w_i = 1/P(R_i = 1)$ for every unit in $\mathcal{C}$, for a sample of size n. Show that

$$E\left[\bar{y}_{\mathcal{C}w} - \bar{y}_{\mathcal{U}}\right] = 0.$$

We use results from Chapter 4 on ratio estimation.

$$\begin{aligned}
E\left[\bar{y}_{\mathcal{C}w}\right] - \bar{y}_{\mathcal{U}} &= \frac{1}{N} E\left[\sum_{i=1}^{N} R_i w_i y_i \left(1 - \frac{\sum_{j=1}^{N} R_j w_j - N}{\sum_{j=1}^{N} R_j w_j}\right)\right] - \bar{y}_{\mathcal{U}} \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{P(R_i = 1)} E\left[R_i\right] - \bar{y}_{\mathcal{U}} - \frac{1}{N} E\left[\sum_{i=1}^{N} R_i w_i y_i \frac{\sum_{j=1}^{N} R_j w_j - N}{\sum_{j=1}^{N} R_j w_j}\right] \\
&= \frac{1}{n} \sum_{i=1}^{N} \frac{y_i}{P(R_i = 1)} P(R_i = 1) \\
&= t_y.
\end{aligned}$$

15.23 When some observations have zero chance of being in the sample. there is no way to guarantee that the estimator of every conceivable population quantity is unbiased and for any estimator of the population mean, it is possible to construct a response variable y for which the estimator is biased. Simply give one of the zero-chance observations an enormous value of y. The only hope is that for the variables of interest, the population units with zero chance of being selected are similar enough to units in the sample that the bias is reduced.

15.26 Answers will vary, but note that the use of a quota sample means that the sample did not have the safeguards for representativeness that would have been afforded by a probability sample. The "initial interviewers were graduate students or postgraduate students in sociology"; if the interviewers had discretion about which persons to interview, results might be skewed in the direction of the interviewers' assumptions.

15.29 All are convenience samples, but (c) has the most scope for manipulation.

There are larger statistical issues, too. Audience members have already passed one test: They were enthusiastic enough about a movie to buy a ticket. That means that people who aren't interested in a movie won't be included at all in the score (unlike professional critics, who watch movies because it's their job and rarely get to pick what they see). Additionally, audience members may not take the time to review the movie if they have neutral feelings about it, which means the range of responses will likely be concentrated toward either the very positive or the very negative.

All those factors mean that the verified audience score will only represent the opinions of people who wanted to see a film enough to spend money, including an online service fee, on the ticket ahead of time (rather than buying at the theater), and who then took the time to log in to review the film. That is not a representative sample of the general public.

15.30 See `https://libguides.rutgers.edu/evidencebasedmedicine` for a medical pyramid. One possible pyramid for survey data might be:

1. Probability samples with no nonresponse or measurement error.
2. Administrative data. This is a census of a population, but it might not be the population you are interested in. These can be considered as either probability or nonprobability samples, depending on what you are interested in.
3. Quota samples.
4. Convenience samples. With convenience samples, the primary characteristic for inclusion in the sample is being easy to get. Examples include stopping people on the street or in a mall, putting up a web site and asking for volunteers,

Chapter 16

Survey Quality

16.1 There could be several sources of measurement error, and student answers will vary for this question. Respondents might not know or remember whether they had signed up on the list. They may be confused about which calls are supposed to be blocked (calls from political candidates are not blocked, for example). Or the question may be worded confusingly.

16.2 One possible source of measurement error might be that people do not remember whether they voted. Or they may say that they voted when they did not because of social desirability. Studies of election behavior often link the survey respondents to their voting record, to see whether they did, in fact, vote. But this can also be problematic. Some of the earlier studies claimed that members of certain groups had a tendency to say they voted when they actually did not, but the problem was likely not that survey respondents were misrepresenting their record, but that the investigators did not link them to the correct voting record.

16.3 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	7.34×10^{-5}
Graduates	1548	150	27	0.1800	0.0245	1.66×10^{-5}
Professional	860	80	27	0.3375	0.0255	1.47×10^{-5}
Total	11380	1130	177		0.1577	1.05×10^{-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.

16.4 Because different months have different numbers of days, to be precise here we have $1 - P = P(\text{birthday in first 10 days}) = 120/365.25 = 0.3285421$. Note that the denominator includes an extra 0.25 to account for leap years. Thus, the probability of answering the sensitive question is $P = 0.6714579$. The probability for the innocuous question is $p_I = 181.25/365.25 = 0.4962355$, although using $p_I = 1/2$ is close enough.

(a) For survey 1, $\hat{\phi} = 551/1202 = 0.4584027$

$$\hat{p}_S = \frac{551/1202 - 0.3285421 * 0.4962355}{0.6714579} = 0.4398912$$

with

$$\hat{V}(\hat{p}_S) = 0.4584027 * (1 - 0.4584027)/(1202 * 0.6714579^2) = 0.0004581225.$$

Thus, a 95% CI for p_S is $0.4398912 \pm 1.96\sqrt{(}0.0004581225) = [0.398, 0.482]$.

(b) For survey 2, $\hat{\phi} = 528/965 = 0.5471503$

$$\hat{p}_S = \frac{0.5471503 - 0.3285421 * 0.4962355}{0.6714579} = 0.5720627$$

with

$$\hat{V}(\hat{p}_S) = 0.5471503 * (1 - 0.5471503)/(965 * 0.6714579^2) = 0.0005695028.$$

Thus, a 95% CI for p_S is $0.5720627 \pm 1.96\sqrt{(}0.0005695028) = [0.525, 0.619]$.

(c) The big assumption is that respondents follow the instructions and answer the questions honestly. A further assumption is made for the variance calculations, that the respondents are independent.

Note that some persons who were approached declined to participate in the survey, so there might be nonresponse bias (although it might be thought that persons who had used doping might be less likely to participate). They also may have misunderstood the format, and just answered the innocuous question.

16.5 (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,

$$\begin{aligned} V(\hat{p}_s) &= \frac{1}{(\theta_1-\theta_2)^2} V(\hat{p}) \\ &= \frac{1}{(\theta_1-\theta_2)^2} \frac{p(1-p)}{n-1}. \end{aligned}$$

16.6 Let μ_i be the original answer of respondent i, and let y_i be the value recorded after the noise has been added.

Let $a_i = 1$ if first coin is heads, with $P(a_i = 1) = P$, and $b_i = 1$ if second coin is heads, with $P(b_i = 1) = p_I$.

$$\begin{aligned} P(y_i = 1|\mu_i = 1) &= P(y_i = 1, a_i = 1|\mu_i = 1) + P(y_i = 1, a_i = 0|\mu_i = 1) \\ &= P(a_i = 1) + P(a_i = 0, b_i = 1) \\ &= P + (1-P)p_I. \end{aligned}$$

$$\begin{aligned} P(y_i = 1|\mu_i = 0) &= P(y_i = 1, a_i = 1|\mu_i = 0) + P(y_i = 1, a_i = 0|\mu_i = 0) \\ &= 0 + P(a_i = 0, b_i = 1) \\ &= (1-P)p_I. \end{aligned}$$

$$\begin{aligned} P(y_i = 0|\mu_i = 0) &= P(y_i = 0, a_i = 1|\mu_i = 0) + P(y_i = 0, a_i = 0|\mu_i = 0) \\ &= P(a_i = 1) + P(a_i = 0, b_i = 0) \\ &= P + (1-P)(1-p_I). \end{aligned}$$

$$\begin{aligned} P(y_i = 0|\mu_i = 1) &= P(y_i = 0, a_i = 1|\mu_i = 1) + P(y_i = 0, a_i = 0|\mu_i = 1) \\ &= 0 + P(a_i = 0, b_i = 0) \\ &= (1 - P)(1 - p_I). \end{aligned}$$

(b)

$$\frac{P(y_i = 1|\mu_i = 1)}{P(y_i = 1|\mu_i = 0)} = \frac{P + (1 - P)p_I}{(1 - P)p_I} = 1 + \frac{P}{(1 - P)p_I},$$

$$\frac{P(y_i = 0|\mu_i = 0)}{P(y_i = 0|\mu_i = 1)} = \frac{P + (1 - P)(1 - p_I)}{(1 - P)(1 - p_I)} = 1 + \frac{P}{(1 - P)(1 - p_I)}.$$

The value M is the larger of these two ratios of probabilities. Note that if $p_I = 1/2$, then each probability equals $1 + 2P/(1 - P)$.

(c) This exploration is best done computationally, although you can also derive theoretical results. I used the R function below. Note that any value for n may be used since this is just a scaling factor.

```
dp = function(pS,pI,P,n=100) {
   phi = pS*P + pI*(1-P)
   vphi = phi*(1-phi)/n
   vpS = vphi/P^2
   M1 = (P + (1-P)*pI)/ ( (1-P)*pI )
   M2 = (P + (1-P)*(1-pI))/ ( (1-P)*(1-pI) )
   M = pmax(M1,M2)
   cbind(vpS,M)
}

pSgrid = seq(.1,.9,.2)
dpgrid = expand.grid(pS = pSgrid,pI = pSgrid, P = pSgrid)
dpout = dp(dpgrid[,1],dpgrid[,2],dpgrid[,3])
```

A choice of $p_I = 0.5$ and $P = 0.5$ appears to work well for most values of p_S.

16.8 The total variability of a response under the model in (16.4) is

$$V(y_{ijt}) = V(\mu) + \sigma_b^2 + \sigma_i^2$$

and

$$\text{Cov}\,(y_{ijt}, y_{ijs}) = \sigma_b^2 \text{if } s \neq t.$$

An ICC for interviewing can be defined as $\text{Cov}\,(y_{ijt}, y_{ijs})/V(y_{ijt})$.

Chapter 17

Appendix A: Probability Concepts Used in Sampling

A.1

$$P\text{ (match exactly 3 numbers)} = \frac{\binom{5}{3}\binom{30}{2}}{\binom{35}{5}} = \frac{(10)(435)}{324{,}632} = \frac{4350}{324{,}632}$$

$$\begin{aligned} P\text{ (match at least 1 number)} &= 1 - P\text{ (match no numbers)} \\ &= 1 - \frac{\binom{5}{0}\binom{30}{5}}{\binom{35}{5}} \\ &= 1 - \frac{142{,}506}{324{,}632} = \frac{182{,}126}{324{,}632}. \end{aligned}$$

A.2

$$P\text{ (no 7s)} \quad = \quad \frac{\binom{3}{0}\binom{5}{4}}{\binom{8}{4}} = \frac{5}{70}$$

$$P \text{ (exactly one 7)} = \frac{\binom{3}{1}\binom{5}{3}}{\binom{8}{4}} = \frac{30}{70}$$

$$P \text{ (exactly two 7s)} = \frac{\binom{3}{2}\binom{5}{2}}{\binom{8}{4}} = \frac{30}{70}.$$

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

$$\begin{aligned} E[XY] &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \\ &= \Big[\sum_x xP(X = x)\Big]\Big[\sum_y yP(Y = y)\Big] \end{aligned}$$

$$= \quad (EX)(EY).$$

Property 4:

$$\begin{aligned}\text{Cov}\,[X,Y] &= E[(X-EX)(Y-EY)] \\ &= E[XY - Y(EX) - X(EY) - (EX)(EY)] \\ &= E[XY] - (EX)(EY).\end{aligned}$$

Property 5: Using Property 4,

$$\begin{aligned}&\text{Cov}\Big[\sum_{i=1}^{n} a_iX_i + b_i, \sum_{j=1}^{m} c_jY_j + d_j\Big] \\ &= E\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}(a_iX_i+b_i)(c_jY_j+d_j)\Big] \\ &\quad -E\Big[\sum_{i=1}^{n}(a_iX_i+b_i)\Big]E\Big[\sum_{j=1}^{m}(c_jY_j+d_j)\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}[a_ic_jE(X_iY_j) + a_id_jEX_i + b_ic_jEY_j + b_id_j] \\ &\quad -\sum_{i=1}^{n}\sum_{j=1}^{m}[a_iE(X_i)+b_i][c_jE(Y_j)+d_j] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_ic_j[E(X_iY_j) - (EX_i)(EY_j)] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_ic_j\text{Cov}\,[X_i,Y_j].\end{aligned}$$

Property 6: Using Property 4,

$$V[X] = \text{Cov}\,(X,X) = E[X^2] - (EX)^2.$$

Property 7: Using Property 5,

$$\begin{aligned}V[X+Y] &= \text{Cov}\,[X+Y, X+Y] \\ &= \text{Cov}\,[X,X] + \text{Cov}\,[Y,X] + \text{Cov}\,[X,Y] + \text{Cov}\,[Y,Y] \\ &= V[X] + V[Y] + 2\text{Cov}\,[X,Y].\end{aligned}$$

(It follows from the definition of Cov that $\text{Cov}\,[X,Y] = \text{Cov}\,[Y,X]$.)

Property 8: From Property 7,

$$\begin{aligned} V\Big[\frac{X}{\sqrt{V(X)}}+\frac{Y}{\sqrt{V(Y)}}\Big] &= V\Big[\frac{X}{\sqrt{V(X)}}\Big]+V\Big[\frac{Y}{\sqrt{V(Y)}}\Big] \\ &\quad +2\text{Cov}\left[\frac{X}{\sqrt{V(X)}},\frac{Y}{\sqrt{V(Y)}}\right] \\ &= 2+2\text{Corr}\,[X,Y]. \end{aligned}$$

Since the variance on the left must be nonnegative, we have $2 + 2\ \text{Corr}\,[X, Y] \geq 0$, which implies $\text{Corr}\,[X, Y] \geq -1$.

Similarly, the relation

$$0 \leq V\Big[\frac{X}{\sqrt{V(X)}}-\frac{Y}{\sqrt{V(Y)}}\Big] = 2-2\ \text{Corr}\,[X,Y]$$

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,

$$\begin{aligned} V[Z_i] &= E[Z_i^2]-[E(Z_i)]^2 \\ &= \frac{n}{N}-\left(\frac{n}{N}\right)^2 \\ &= \frac{n(N-n)}{N^2}. \end{aligned}$$

$$\begin{aligned} \text{Cov}\,[Z_i,Z_j] &= E[Z_iZ_j]-(EZ_i)(EZ_j) \\ &= \frac{n(n-1)}{N(N-1)}-\left(\frac{n}{N}\right)^2 \\ &= \frac{n(n-1)N-n^2(N-1)}{N^2(N-1)} \\ &= -\frac{n(N-n)}{N^2(N-1)}. \end{aligned}$$

A.5

$$\begin{aligned} \text{Corr}\,[\bar{x},\bar{y}] &= \frac{\text{Cov}\,[\bar{x},\bar{y}]}{\sqrt{V[\bar{x}]V[\bar{y}]}} \\ &= \frac{\frac{1}{n}(1-\frac{n}{N})RS_xS_y}{\sqrt{[\frac{1}{n}(1-\frac{n}{N})S_x^2][\frac{1}{n}(1-\frac{n}{N})S_y^2]}} \\ &= R. \end{aligned}$$

A.6 We can find the three conditional expectations of Y and Y^2:

$$\begin{aligned}
E[Y \mid Z = A] &= \frac{4+9+7+6}{4} = \frac{13}{2} \\
E[Y \mid Z = B] &= \frac{6+8}{2} = 7 \\
E[Y \mid Z = C] &= 5 \\
E[Y^2 \mid Z = A] &= \frac{4^2+9^2+7^2+6^2}{4} = \frac{182}{4} = \frac{91}{2} \\
E[Y^2 \mid Z = B] &= \frac{6^2+8^2}{2} = 50 \\
E[Y^2 \mid Z = C] &= 25
\end{aligned}$$

Then,

$$\begin{aligned}
E[Y] &= \frac{1}{3}\left[\frac{13}{2} + 7 + 5\right] = \frac{37}{6} \\
E[Y^2] &= \frac{1}{3}\left[\frac{91}{2} + 50 + 25\right] = \frac{241}{6}
\end{aligned}$$

and

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{241}{6} - \left(\frac{37}{6}\right)^2.$$

Bibliography

Alf, C. and S. L. Lohr (2007). Sampling assumptions in introductory statistics classes. *The American Statistician 61*, 71–77.

Christensen, R. (2020). *Plane Answers to Complex Questions, Fifth Edition*. New York: Springer.

Cochran, W. G. (1977). *Sampling Techniques, 3rd ed.* New York: Wiley.

Demont, P. (2010). Allotment and democracy in ancient Greece. `https://booksandideas.net/Allotment-and-Democracy-in-Ancient.html` (accessed August 10, 2020).

Dvorak, P. (2011, December 22). What is the point of having kids if your life ends when theirs begins? *The Washington Post*, `https://www.washingtonpost.com/local/what--is--the--point--of--having--kids--if--your--life--ends--when--theirs--begins/2011/12/22/gIQAIVkeBP_story.html` (accessed January 30, 2020).

Finucan, H., R. Galbraith, and M. Stone (1974). Moments without tears in simple random sampling from a finite population. *Biometrika 61*, 151–154.

Frankovic, K. (2008). Race, gender and bias in the electorate. `www.cbsnews.com/stories/2008/03/17/opinions/pollpositons` (accessed August 15, 2008).

Fryar, C. D., D. Kruszon-Moran, Q. Gu, and C. L. Ogden (2018). Mean body weight, height, waist circumference, and body mass index among adults: United States, 1999-2000 through 2015-2016. *National Health Statistics Reports* (122).

Geldsetzer, P., G. Fink, M. Vaikath, and T. Bärnighausen (2018). Sampling for patient exit interviews: assessment of methods using mathematical derivation and computer simulations. *Health Services Research 53*(1), 256–272.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953). *Sample Survey Methods and Theory. Volume 2: Theory*. New York: Wiley.

Kinsley, M. (1981). The art of polling. *New Republic 184*, 16–19.

Landers, A. (1976). If you had it to do over again, would you have children? *Good Housekeeping 182* (June), 100–101, 215–216, 223–224.

Landers, A. (1996). *Wake Up and Smell the Coffee!* New York: Villard.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York: Springer-Verlag.

Lohr, S. L. (2008). Coverage and sampling. In E. deLeeuw, J. Hox, and D. Dillman (Eds.), *International Handbook of Survey Methodology*, pp. 97–112. New York: Erlbaum.

Lohr, S. L. (2022). *SAS® Software Companion for* Sampling: Design and Analysis, 3rd ed. Boca Raton, FL: CRC Press.

Lu, Y. and S. L. Lohr (2022). *R Companion for* Sampling: Design and Analysis, 3rd ed. Boca Raton, FL: CRC Press.

Mccallum, M. L. and G. W. Bury (2013). Google search patterns suggest declining interest in the environment. *Biodiversity and Conservation 22*(6–7), 1355–1367.

Mez, J., D. H. Daneshvar, P. T. Kiernan, B. Abdolmohammadi, V. E. Alvarez, B. R. Huber, M. L. Alosco, T. M. Solomon, C. J. Nowinski, L. McHale, et al. (2017). Clinicopathological evaluation of chronic traumatic encephalopathy in players of American football. *Journal of the American Medical Association 318*(4), 360–370.

Newport, F. and J. Wilke (2013). Desire for children still norm in U.S. `https://news.gallup.com/poll/164618/desire-children-norm.aspx` (accessed January 31, 2020).

Raj, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika 60*, 125–133.

Stasko, E. C. (2015, July). Sexting and intimate partner relationships among adults. Master's thesis, Drexel University, Philadelphia. `https://idea.library.drexel.edu/islandora/object/idea%3A6558/datastream/OBJ/download/Sexting_and_Intimate_Partner_Relationships_among_Adults.pdf` (accessed October 23, 2019).

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (2014, June). *National Intimate Partner and Sexual Violence Survey (NISVS): General Population Survey Raw Data, 2010: User's Guide*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor].

Zhang, C., C. Antoun, H. Y. Yan, and F. G. Conrad (2020). Professional respondents in opt-in online panels: What do we really know? *Social Science Computer Review 38*(6), 703–719.