Chapter 1

Introduction

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

Sampling frame: Persons who know about the telephone survey.

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

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

1.2 Target population: All mutual funds.

Sampling frame: Mutual funds listed in newspaper.

Sampling unit = observation unit: One listing.

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

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

Sampling frame: Persons who visit the website

Sampling unit = observation unit: One review.

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

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

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

Sampling unit = observation unit: One resident.

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

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

1.5 Target population: All homeless persons in study area.

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

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

Observation unit: Person.

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

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

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

Sampling unit = observation unit: One woman.

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

1.7 Target population: All cows in region.

Sampling frame: List of all farms in region.

Sampling unit: One farm.

Observation unit: One cow.

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

stage cluster sample, discussed in Chapter 5.

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

Sampling frame: List of 184 licensed homes.

Sampling unit = observation unit: One home.

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

1.13 Target population: All attendees of the 2005 JSM.

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

Sampling unit: One e-mail address.

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

1.14 Target population: All professors of education

Sampling population: List of education professors

Sampling unit: One professor

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

1.15 Target population: All adults

Sampling population: Friends and relatives of American Cancer Society volunteers

Sampling unit: One person

Here's what I wrote about the survey elsewhere:

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

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

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

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

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

Chapter 2

Simple Probability Samples

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

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

Plan 1:

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

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

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

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

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

Plan 2:

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

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

(ii)

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

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

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

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

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

Similarly,

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

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

(b)

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

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

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

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

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

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

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