Chapter 1

Introduction

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

Sampling frame: Persons who know about the telephone survey.

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

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

1.2 Target population: All mutual funds.

Sampling frame: Mutual funds listed in newspaper.

Sampling unit = observation unit: One listing.

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

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

Sampling frame: Persons who visit the website

Sampling unit = observation unit: One review.

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

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

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

Sampling unit = observation unit: One resident.

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

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

1.5 Target population: All homeless persons in study area.

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

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

Observation unit: Person.

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

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

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

Sampling unit = observation unit: One woman.

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

1.7 Target population: All cows in region.

Sampling frame: List of all farms in region.

Sampling unit: One farm.

Observation unit: One cow.

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

stage cluster sample, discussed in Chapter 5.

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

Sampling frame: List of 184 licensed homes.

Sampling unit = observation unit: One home.

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

1.13 Target population: All attendees of the 2005 JSM.

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

Sampling unit: One e-mail address.

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

1.14 Target population: All professors of education

Sampling population: List of education professors

Sampling unit: One professor

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

1.15 Target population: All adults

Sampling population: Friends and relatives of American Cancer Society volunteers

Sampling unit: One person

Here's what I wrote about the survey elsewhere:

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

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

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

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

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

Chapter 2

Simple Probability Samples

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

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

Plan 1:

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

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

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

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

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

Plan 2:

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

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

(ii)

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

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

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

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

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

Similarly,

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

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

(b)

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

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

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

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

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

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

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

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

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

Using the sampling distribution,

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

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

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

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

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

(b) A total of $8^3 = 512$ samples are possible when sampling with replacement. Fortunately, we need not list all of these to find the sampling distribution of $\bar{y}$. Let X_i be the value of the 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 2.28),

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

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

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

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

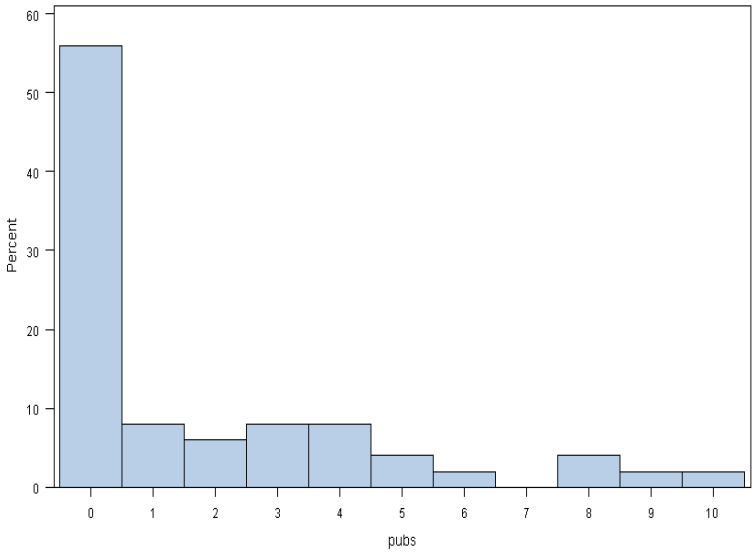
$$\text{SE}\,(\hat{t}) = \sqrt{3728.238} = 61.0593$$

and a 95% CI for t is

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

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

2.6 (a)



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

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

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

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

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

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

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

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

observations to use the central limit theorem.

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

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

A 95% confidence interval is

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

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

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

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

2.08 Answers will vary.

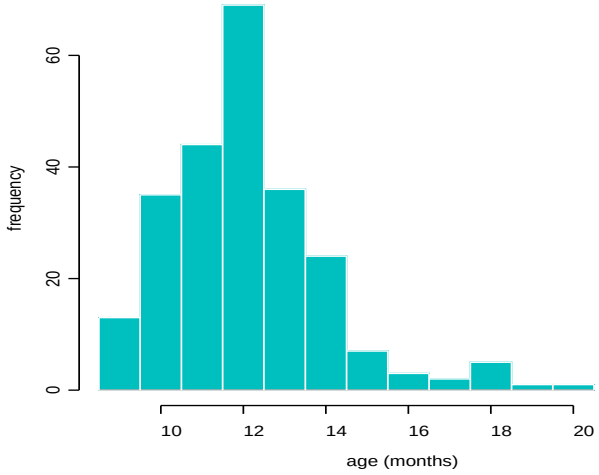
2.09 If $n_0 \leq N$, then

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

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

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

2.11 (a)



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

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

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

A 95% confidence interval is

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

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

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

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

Then

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

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

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

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

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

so an approximate 95% CI is

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

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

2.14 Here is SAS output:

```
The SURVEYMEANS Procedure

            Data Summary

Number of Observations           150
Sum of Weights                   864

   Class Level Information

Class
Variable      Levels    Values

sex                2    f m


                          Statistics

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

                          Statistics

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

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

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

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

CI : $[556, 642]$

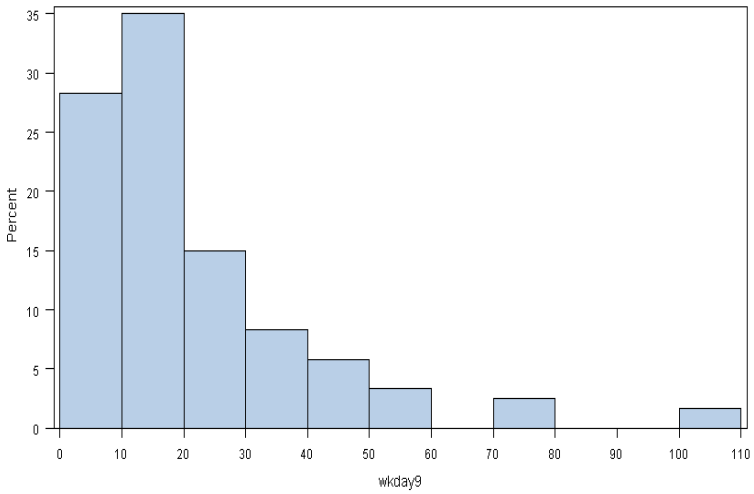
(c) $\bar{y} = 56.593$, $s^2 = 5292.73$

CI : $[48.8, 64.4]$

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

CI : $[39.7, 54.0]$

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



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

Here is SAS code for problems 2.16 and 2.17:

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

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

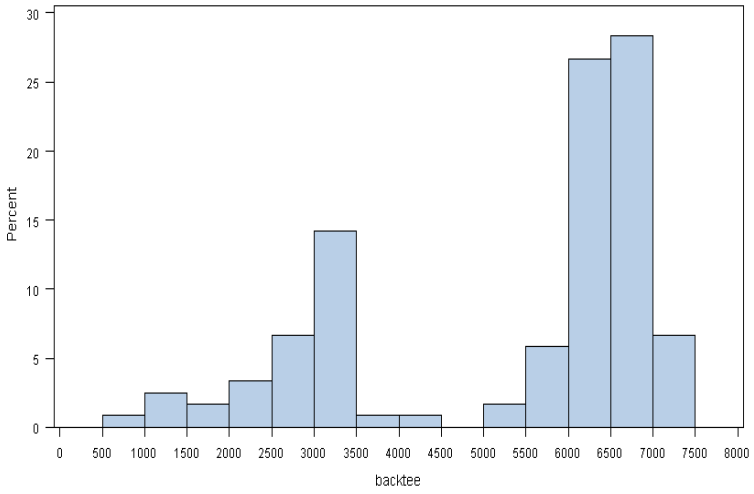
```
   input  RN  state $ holes  type $ yearblt  wkday18  wkday9  wkend18
      wkend9  backtee  rating  par  cart18  cart9  caddy $ pro  $ ;

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

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

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

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



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

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

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

or [0.627, 0.790].

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

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

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

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

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

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

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

and

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

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

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

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

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

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

so the procedure results in a simple random sample.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.22 (a) From (2.13),

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

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

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

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

(b) I used Excel to calculate these values.

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

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

2.23

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

2.24

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

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

Setting the derivative equal to 0 and solving for n gives

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

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

2.25 (a) Skewed, with tail on right.

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

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

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

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

2.27 (a)

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

(b)

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

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

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

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

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

and

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

(b)

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

(c)

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

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

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

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

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

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

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

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

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