assumption is that of nonresponse. Response levels of only 60% to 70% are common in surveys of human populations. The possibility of substantial bias clearly arises if there is a relationship of potential answers to survey questions to the propensity to respond to those questions. For example, adults living in families are easier to contact by a telephone survey than those living alone, and the opinions of these two groups may well differ on certain issues. It is important to realize that the standard errors of estimates that we have developed earlier in this chapter account only for random variability in sample composition, not for systematic biases.

The *Literary Digest* poll of 1936, which predicted a 57% to 43% victory for Republican Alfred Landon over incumbent president Franklin Roosevelt, is one of the most famous of flawed surveys. Questionnaires were mailed to about 10 million voters, who were selected from lists such as telephone books and club memberships, and approximately 2.4 million of the questionnaires were returned. There were two intrinsic problems: (1) nonresponse—those who did not respond may have voted differently from those who did—and (2) selection bias—even if all 10 million voters had responded, they would not have constituted a random sample; those in lower socioeconomic classes (who were more likely to vote for Roosevelt) were less likely to have telephone service or belong to clubs and thus less likely to be included in the sample than were wealthier voters. The assumption that an exact measurement is obtained from every member of the sample may also be in error. In surveys conducted by interviewers, the interviewer's approach and personality may affect the response. In surveys that use questionnaires, the wording of the questions and the context within which they are lodged can have an effect. An interesting example is a poll conducted by Stanley Presser, (*New Yorker*, Oct 18, 2004). Half of the sample was asked, "Do you think the United States should allow public speeches against democracy?" The other half was asked, "Do you think the United States should forbid public speeches against democracy?" 56% said no to the first question, and 39% said yes to the second. The interesting paper by Hansen in Tanur et al. (1972) reports on efforts of the U.S. Bureau of the Census to investigate these sorts of problems.

7.7 Problems

1. Consider a population consisting of five values—1, 2, 2, 4, and 8. Find the population mean and variance. Calculate the sampling distribution of the mean of a sample of size 2 by generating all possible such samples. From them, find the mean and variance of the sampling distribution, and compare the results to Theorems A and B in Section 7.3.1.

2. Suppose that a sample of size $n = 2$ is drawn from the population of the preceding problem and that the proportion of the sample values that are greater than 3 is recorded. Find the sampling distribution of this statistic by listing all possible such samples. Find the mean and variance of the sampling distribution.

3. Which of the following is a random variable?

a. The population mean

b. The population size, N

c. The sample size, n
d. The sample mean
e. The variance of the sample mean
f. The largest value in the sample
g. The population variance
h. The estimated variance of the sample mean

4. Two populations are surveyed with simple random samples. A sample of size n_1 is used for population I, which has a population standard deviation σ_1; a sample of size $n_2 = 2n_1$ is used for population II, which has a population standard deviation $\sigma_2 = 2\sigma_1$. Ignoring finite population corrections, in which of the two samples would you expect the estimate of the population mean to be more accurate?

5. How would you respond to a friend who asks you, "How can we say that the sample mean is a random variable when it is just a number, like the population mean? For example, in Example A of Section 7.3.2, a simple random sample of size 50 produced $\bar{x} = 938.5$; how can the number 938.5 be a random variable?"

6. Suppose that two populations have equal population variances but are of different sizes: $N_1 = 100{,}000$ and $N_2 = 10{,}000{,}000$. Compare the variances of the sample means for a sample of size $n = 25$. Is it substantially easier to estimate the mean of the smaller population?

7. Suppose that a simple random sample is used to estimate the proportion of families in a certain area that are living below the poverty level. If this proportion is roughly .15, what sample size is necessary so that the standard error of the estimate is .02?

8. A sample of size $n = 100$ is taken from a population that has a proportion $p = 1/5$.

a. Find δ such that $P(|\hat{p} - p| \geq \delta) = 0.025$.
b. If, in the sample, $\hat{p} = 0.25$, will the 95% confidence interval for p contain the true value of p?

9. In a simple random sample of 1,500 voters, 55% said they planned to vote for a particular proposition, and 45% said they planned to vote against it. The estimated margin of victory for the proposition is thus 10%. What is the standard error of this estimated margin? What is an approximate 95% confidence interval for the margin?

10. True or false (and state why):
If a sample from a population is large, a histogram of the values in the sample will be approximately normal, even if the population is not normal.

11. Consider a population of size four, the members of which have values x_1, x_2, x_3, x_4.

a. If simple random sampling were used, how many samples of size two are there?
b. Suppose that rather than simple random sampling, the following sampling scheme is used. The possible samples of size two are

$$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\}, \{x_1, x_4\}$$

and the sampling is done in such a way that each of these four possible samples is equally likely. Is the sample mean unbiased?

12. Consider simple random sampling *with* replacement.

a. Show that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimate of σ^2.

b. Is s an unbiased estimate of σ?

c. Show that $n^{-1}s^2$ is an unbiased estimate of $\sigma^2_{\overline{X}}$.

d. Show that $n^{-1}N^2s^2$ is an unbiased estimate of σ^2_T.

e. Show that $\hat{p}(1-\hat{p})/(n-1)$ is an unbiased estimate of $\sigma^2_{\hat{p}}$.

13. Suppose that the total number of discharges, τ, in Example A of Section 7.2 is estimated from a simple random sample of size 50. Denoting the estimate by T, use the central limit theorem to sketch the approximate probability density of the error $T - \tau$.

14. The proportion of hospitals in Example A of Section 7.2 that had fewer than 1000 discharges is $p = .654$. Suppose that the total number of hospitals having fewer than 1000 discharges is estimated from a simple random sample of size 25. Use the central limit theorem to sketch the approximate sampling distribution of the estimate.

15. Consider estimating the mean of the population of hospital discharges (Example A of Section 7.2) from a simple random sample of size n. Use the normal approximation to the distribution of $\overline{X}$ in answering the following:

a. Sketch $P(|\overline{X} - \mu| > 200)$ as a function of n for $20 \le n \le 100$.

b. For $n = 20, 40$, and 80, find Δ such that $P(|\overline{X} - \mu| > \Delta) \approx .10$. Similarly, find Δ such that $P(|\overline{X} - \mu| > \Delta) \approx .50$.

16. True or false?

a. The center of a 95% confidence interval for the population mean is a random variable.

b. A 95% confidence interval for μ contains the sample mean with probability .95.

c. A 95% confidence interval contains 95% of the population.

d. Out of one hundred 95% confidence intervals for μ, 95 will contain μ.

17. A 90% confidence interval for the average number of children per household based on a simple random sample is found to be (.7, 2.1). Can we conclude that 90% of households have between .7 and 2.1 children?

18. From independent surveys of two populations, 90% confidence intervals for the population means are constructed. What is the probability that neither interval contains the respective population mean? That both do?

19. This problem introduces the concept of a *one-sided confidence interval.* Using the central limit theorem, how should the constant k be chosen so that the interval

$(-\infty, \overline{X} + ks_{\overline{X}})$ is a 90% confidence interval for μ—i.e., so that $P(\mu \leq \overline{X} + ks_{\overline{X}}) = .9$? This is called a one-sided confidence interval. How should k be chosen so that $(\overline{X} - ks_{\overline{X}}, \infty)$ is 95% one-sided confidence interval?

20. In Example D of Section 7.3.3, a 95% confidence interval for μ was found to be $(1.44, 1.76)$. Because μ is some fixed number, it either lies in this interval or it doesn't, so it doesn't make any sense to claim that $P(1.44 \leq \mu \leq 1.76) = .95$. What do we mean, then, by saying this is a "95% confidence interval?"

21. In order to halve the width of a 95% confidence interval for a mean, by what factor should the sample size be increased? Ignore the finite population correction.

22. An investigator quantifies her uncertainty about the estimate of a population mean by reporting $\overline{X} \pm s_{\overline{X}}$. What size confidence interval is this?

23. **a.** Show that the standard error of an estimated proportion is largest when $p = 1/2$.

b. Use this result and Corollary B of Section 7.3.2 to conclude that the quantity

$$\frac{1}{2}\sqrt{\frac{N-n}{N(n-1)}}$$

is a conservative estimate of the standard error of $\hat{p}$ no matter what the value of p may be.

c. Use the central limit theorem to conclude that the interval

$$\hat{p} \pm \sqrt{\frac{N-n}{N(n-1)}}$$

contains p with probability at least .95.

24. For a random sample of size n from a population of size N, consider the following as an estimate of μ:

$$\overline{X}_c = \sum_{i=1}^{n} c_i X_i$$

where the c_i are fixed numbers and $X_1, \ldots, X_n$ is the sample.

a. Find a condition on the c_i such that the estimate is unbiased.

b. Show that the choice of c_i that minimizes the variances of the estimate subject to this condition is $c_i = 1/n$, where $i = 1, \ldots, n$.

25. Here is an alternative proof of Lemma B in Section 7.3.1. Consider a random permutation $Y_1, Y_2, \ldots, Y_N$ of $x_1, x_2, \ldots, x_N$. Argue that the joint distribution of any subcollection, $Y_{i_1}, \ldots, Y_{i_n}$, of the Y_i is the same as that of a simple random sample, $X_1, \ldots, X_n$. In particular,

$$\text{Var}(Y_i) = \text{Var}(X_k) = \sigma^2$$

and

$$\text{Cov}(Y_i, Y_j) = \text{Cov}(X_k, X_l) = \gamma$$

if $i \neq j$ and $k \neq l$. Since $Y_1 + Y_2 + \cdots + Y_N = \tau$,

$$\mathrm{Var}\left(\sum_{i=1}^{N} Y_i\right) = 0$$

(Why?) Express $\mathrm{Var}(\sum_{i=1}^{N} Y_i)$ in terms of σ^2 and the unknown covariance, γ. Solve for γ, and conclude that

$$\gamma = -\frac{\sigma^2}{N-1}$$

for $i \neq j$.

26. This is another proof of Lemma B in Section 7.3.1. Let U_i be a random variable with $U_i = 1$ if the ith population member is in the sample and equal to 0 otherwise.

a. Show that the sample mean $\overline{X} = n^{-1} \sum_{i=1}^{N} U_i x_i$.
b. Show that $P(U_i = 1) = n/N$. Find $E(U_i)$, using the fact that U_i is a Bernoulli random variable.
c. What is the variance of the Bernoulli random variable U_i?
d. Noting that $U_i U_j$ is a Bernoulli random variable, find $E(U_i U_j)$, $i \neq j$. (Be careful to take into account that the sample is drawn without replacement.)
e. Find $\mathrm{Cov}(U_i, U_j)$, $i \neq j$.
f. Using the representation of $\overline{X}$ above, find $\mathrm{Var}(\overline{X})$.

27. Suppose that the population size N is not known, but it is known that $n \leq N$. Show that the following procedure will generate a simple random sample of size n. Imagine that the population is arranged in a long list that you can read sequentially.

a. Let the sample initially consist of the the first n elements in the list.
b. For $k = 1, 2, \ldots,$ as long as the end of the list has not been encountered:

i. Read the $(n + k)$-th element in the list.
ii. Place it in the sample with probability $n/(n + k)$ and, if it is placed in the sample, randomly drop one of the exisiting sample members.

28. In surveys, it is difficult to obtain accurate answers to sensitive questions such as "Have you ever used heroin?" or "Have you ever cheated on an exam?" Warner (1965) introduced the method of **randomized response** to deal with such situations. A respondent spins an arrow on a wheel or draws a ball from an urn containing balls of two colors to determine which of two statements to respond to: (1) "I have characteristic A," or (2) "I do not have characteristic A." The interviewer does not know which statement is being responded to but merely records a yes or a no. The hope is that an interviewee is more likely to answer truthfully if he or she realizes that the interviewer does not know which statement is being responded to. Let R be the proportion of a sample answering Yes. Let p be the probability that statement 1 is responded to (p is known from the structure of the randomizing device), and let q be the proportion of the population that has characteristic A. Let r be the probability that a respondent answers Yes.

a. Show that $r = (2p-1)q + (1-p)$. [*Hint:* $P(\text{yes}) = P(\text{yes given question 1}) \times P(\text{question 1}) + P(\text{yes given question 2}) \times P(\text{question 2})$.]

b. If r were known, how could q be determined?

c. Show that $E(R) = r$, and propose an estimate, Q, for q. Show that the estimate is unbiased.

d. Ignoring the finite population correction, show that

$$\text{Var}(R) = \frac{r(1-r)}{n}$$

where n is the sample size.

e. Find an expression for $\text{Var}(Q)$.

29. A variation of the method described in Problem 28 has been proposed. Instead of responding to statement 2, the respondent answers an unrelated question for which the probability of a "yes" response is known, for example, "Were you born in June?"

a. Propose an estimate of q for this method.

b. Show that the estimate is unbiased.

c. Obtain an expression for the variance of the estimate.

30. Compare the accuracies of the methods of Problems 28 and 29 by comparing their standard deviations. You may do this by substituting some plausible numerical values for p and q.

31. Referring to Example D in Section 7.3.3, how large should the sample be in order that the 95% confidence interval for the total number of owners planning to sell will have a width of 500?

32. Referring again to Example D in Section 7.3.3, suppose that a survey is done of another condominium project of 12,000 units. The sample size is 200, and the proportion planning to sell in this sample is .18.

a. What is the standard error of this estimate? Give a 90% confidence interval.

b. Suppose we use the notation $\hat{p}_1 = .12$ and $\hat{p}_2 = .18$ to refer to the proportions in the two samples. Let $\hat{d} = \hat{p}_1 - \hat{p}_2$ be an estimate of the difference, d, of the two population proportions p_1 and p_2. Using the fact that $\hat{p}_1$ and $\hat{p}_2$ are independent random variables, find expressions for the variance and standard error of $\hat{d}$.

c. Because $\hat{p}_1$ and $\hat{p}_2$ are approximately normally distributed, so is $\hat{d}$. Use this fact to construct 99%, 95%, and 90% confidence intervals for d. Is there clear evidence that p_1 is really different from p_2?

33. Two populations are independently surveyed using simple random samples of size n, and two proportions, p_1 and p_2, are estimated. It is expected that both population proportions are close to .5. What should the sample size be so that the standard error of the difference, $\hat{p}_1 - \hat{p}_2$, will be less than .02?

34. In a survey of a very large population, the incidences of two health problems are to be estimated from the same sample. It is expected that the first problem will affect about 3% of the population and the second about 40%. Ignore the finite population correction in answering the following questions.

the estimate

›sed. Instead
question for
ere you born

ıparing their
le numerical

e be in order
nning to sell

:y is done of
200, and the

ce interval.
proportions
:rence, d, of
$_1$ and $\hat{p}_2$ are
ınd standard

$\hat{d}$. Use this
s there clear

samples of
:d that both
e so that the

roblems are
roblem will
re the finite

a. How large should the sample be in order for the standard errors of both estimates to be less than .01? What are the actual standard errors for this sample size?

b. Suppose that instead of imposing the same limit on both standard errors, the investigator wants the standard error to be less than 10% of the true value in each case. What should the sample size be?

35. A simple random sample of a population of size 2000 yields the following 25 values:

104	109	111	109	87
86	80	119	88	122
91	103	99	108	96
104	98	98	83	107
79	87	94	92	97

a. Calculate an unbiased estimate of the population mean.

b. Calculate unbiased estimates of the population variance and Var$(\overline{X})$.

c. Give approximate 95% confidence intervals for the population mean and total.

36. With simple random sampling, is $\overline{X}^2$ an unbiased estimate of μ^2? If not, what is the bias?

37. Two surveys were independently conducted to estimate a population mean, μ. Denote the estimates and their standard errors by $\overline{X}_1$ and $\overline{X}_2$ and $\sigma_{\overline{X}_1}$ and $\sigma_{\overline{X}_2}$. Assume that $\overline{X}_1$ and $\overline{X}_2$ are unbiased. For some α and β, the two estimates can be combined to give a better estimator:

$$X = \alpha\overline{X}_1 + \beta\overline{X}_2$$

a. Find the conditions on α and β that make the combined estimate unbiased.

b. What choice of α and β minimizes the variances, subject to the condition of unbiasedness?

38. Let $X_1, \ldots, X_n$ be a simple random sample. Show that $\frac{1}{n}\sum_{i=1}^{n} X_i^3$ is an unbiased estimate of $\frac{1}{N}\sum_{i=1}^{N} x_i^3$.

39. Suppose that of a population of N items, k are defective in some way. For example, the items might be documents, a small proportion of which are fraudulent. How large should a sample be so that with a specified probability it will contain at least one of the defective items? For example, if $N = 10{,}000$, $k = 50$, and $p = .95$, what should the sample size be? Such calculations are useful in planning sample sizes for acceptance sampling.

40. This problem presents an algorithm for drawing a simple random sample from a population in a sequential manner. The members of the population are considered for inclusion in the sample one at a time in some prespecified order (for example, the order in which they are listed). The ith member of the population is included

61. a. $\mathrm{Cov}(x, Y) = 1/36$; $\mathrm{Corr}(X, Y) = 1/2$
b. $E(X|Y) = Y/2$, $E(Y|X) = (X + 1)/2$
c. If $Z = E(X|Y)$, the density of Z is $f_Z(z) = 8z$, $0 \le z \le 1/2$
If $Z = E(Y|X)$, the density of Z is $f_Z(z) = 8(1 - z)$, $1/2 \le z \le 1$
d. $\hat{Y} = \frac{1}{2} + \frac{1}{2}X$; the mean squared prediction error is 1/24
e. $\hat{Y} = \frac{1}{2} + \frac{1}{2}X$; the mean squared prediction error is 1/24

63. a. $\mathrm{Cov}(X, Y) = -.0085$; $\rho_{XY} = -.1256$
b. $E(Y|X) = (6X^2 + 8X + 3)/[4(3X^2 + 3X + 1)]$

65. In the claim that $E(T\,|N = n) = nE(X)$ **67.** 3/2, 1/6

71. $p_{Y|X}(y\,|x)$ is hypergeometric. $E(Y|X = x) = mx/n$

73. $np(1 + p)$ **75. a.** $1/2\lambda$; **b.** $5/12\lambda^2$

77. $E(X|Y) = Y/2$, $E(Y|X) = X + 1$

79. $M(t) = \frac{1}{2} + \frac{3}{8}e^t + \frac{1}{8}e^{2t}$ **81.** $M(t) = 1 - p + pe^t$

85. $M(t) = e^t p/[1 - (1 - p)e^t]$; $E(X) = 1/p$; $\mathrm{Var}(X) = (1 - p)/p^2$

87. Same p **93.** Exponential

99. b. $E[g(X)] \approx \log\mu - \sigma^2/2\mu^2$; $\mathrm{Var}[g(X)] \approx \sigma^2/\mu^2$

101. $E(Y) \approx \sqrt{\lambda} - 1/(8\sqrt{\lambda})$; $\mathrm{Var}(Y) \approx 1/4$ **103.** .0628 mm

Chapter 5

3. .0228 **13.** $N(0,\ 150{,}000)$; most likely to be where he started

15. $p = .017$ **17.** $n = 96$

21. b. $\mathrm{Var}(\hat{I}(f)) = \dfrac{1}{n}\left[\displaystyle\int_a^b \frac{f^2(x)}{g(x)}dx - I^2(f)\right]$

29. Let $Z_n = n(U_{(n)} - 1)$. Then $P(Z_n \le z) \to e^z$, $-1 \le z \le 0$

Chapter 6

3. $c = .17$ **9.** $E(S^2) = \sigma^2$; $\mathrm{Var}(S^2) = 2\sigma^4/(n - 1)$

Chapter 7

1. $p(1.5) = 1/5$, $p(2) = 1/10$, $p(2.5) = 1/10$, $p(3) = 1/5$, $p(4.5) = 1/10$, $p(5) = 1/5$, $p(6) = 1/10$; $E(\overline{X}) = 17/5$; $\mathrm{Var}(\overline{X}) = 2.34$

3. d, f, h **7.** $n = 319$, ignoring the fpc

9. SE = .026. CI: (.05, .15) **11. a.** 6 samples. **b.** Yes

15. b.

n	Δ_1	Δ_2
20	211.6	86.8
40	145.6	59.7
80	96.9	39.8

17. no

19. 1.28, 1.645

21. The sample size should be multiplied by 4.

29. a. $\hat{Q} = \dfrac{R - t(1-p)}{p}$, where t = probability of answering yes to unrelated question

c. $\text{Var}(\hat{Q}) = r(1-r)/(np^2)$, where $r = P(yes) = qp + t(1-p)$

31. $n = 395$

33. The sample size for each survey should be 1250.

35. a. $\overline{X} = 98.04$

b. $s^2\dfrac{N-1}{N} = 133.64$, $\dfrac{s^2}{n}\left(1 - \dfrac{n}{N}\right) = 5.28$

c. 98.04 ± 4.50 and $196{,}080 \pm 9008$

37. a. $\alpha + \beta = 1$

b. $\alpha = \dfrac{\sigma^2_{\overline{X}_2}}{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}}$ $\quad \beta = \dfrac{\sigma^2_{\overline{X}_1}}{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}}$

39. Choose n such that $p = 1 - \dfrac{(N-k)(N-k-1)\cdots(N-n+k-1)}{N(N-1)\cdots(N-k+1)}$, which can be done by a recursive computation; $n = 581$

41. b. $\dfrac{N^2}{n}\left(\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B\right)$

c. The proposed method has smaller variance if $\rho > \dfrac{\sigma_B^2}{2\sigma_A\sigma_B}$.

d. The ratio estimate is biased. The approximate variance of the ratio estimate is greater if $\dfrac{\mu_A}{\mu_B} > 1$.

43. $R = \dfrac{\overline{V}}{\overline{O}} = .73$, $s_R = .02$, $.73 \pm .04$

47. The bias is .96 for $n = 64$ and .39 for $n = 128$.

48 **49.** Ignoring the fpc,

a. $R = 31.25$;

b. $s_R = .835$; 31.25 ± 1.637;

c. $T = 10^7$; $10^7 \pm 5, 228, 153$;

d. $s_{T_R} = 266{,}400$, which is much better.

53. a. For optimal allocation, the sample sizes are 10, 18, 17, 19, 12, 9, 15. For proportional allocation they are 20, 23, 19, 17, 8, 6, 7.

b. $\text{Var}(\overline{X}_{SO}) = 2.90$, $\text{Var}(\overline{X}_{sp}) = 3.4$, $\text{Var}(\overline{X}_{srs}) = 6.2$