CHAPTER 4

Proportions, Percentages, and Counts

4.1 Introduction

Employment rates, educational levels, economic status, and health conditions of the public are estimated at regular intervals through the surveys of national organizations. Marketing analysts estimate preferences of the public from the surveys on consumer panels. Private agencies obtain opinions of the public on issues such as economic situation, school budgets, health care programs, and changes in legislation. Polls are conducted to ascertain the choices of the public on political appointments and to predict outcomes of elections.

To estimate the proportions or percentages and the total numbers or counts of the population in the above types of categories or classes, respondents may be asked to provide a simple "Yes" or "No" answer to each of the survey questions. Similarly, responses to some types of questions may take the form of *approve* or *disapprove* and *agree* or *disagree*. Some surveys related to opinions and attitudes may request the respondents to specify their preferences on a scale usually ranging from zero or one to five or ten. The options *Uncertain*, *Undecided*, or *Don't know* are added to some questions in opinion surveys.

In this chapter, estimation of the population proportions and total counts for the case of two classes is described first, and then extended to three or more classes. Topics in the following sections also include confidence limits for the proportions and total counts, methods for finding sample sizes for specified criteria, and estimation of population sizes.

4.2 Two classes

Consider a qualitative characteristic with the Yes–No classification. Let C denote the number of the N population units falling into the first class. The proportion of the units in this class is

$$P = \frac{C}{N}. \tag{4.1}$$

The complementary group consists of $(N - C)$ units and their proportion is $Q = (N - C)/N = 1 - P$.

In a random sample of n units selected without replacement from the N units, denote the numbers of units falling into the two classes by c and $(n - c)$. The sample proportion of the units in the first group is

$$p = \frac{c}{n}. \tag{4.2}$$

The proportion in the complementary group is $q = (n - \mathrm{c})/n = 1 - p$.

Correspondence among the mean, total, and variance of a quantitative variable considered in Chapter 2 and the proportion and count of a qualitative characteristic can be established as follows. Assigning the numerals 1 or 0 to y_i if a unit belongs to the first or the second group, the total, mean, and variance become

$$Y = \sum_{1}^{N} y_i = C, \qquad \bar{Y} = \frac{C}{N} = P, \tag{4.3}$$

and

$$S^2 = \frac{\sum^{N} y_i^2 - N\bar{Y}^2}{N-1} = \frac{C - NP^2}{N-1} = \frac{N}{N-1}PQ. \tag{4.4}$$

Similarly, with this **mnemonic** notation,

$$\bar{y} = \frac{\sum_{1}^{n} y_i}{n} = \frac{c}{n} = p \tag{4.5}$$

and

$$s^2 = \frac{\sum_{1}^{N} y_i^2 - n\bar{y}^2}{n-1} = \frac{c - np^2}{n-1} = \frac{n}{n-1}pq. \tag{4.6}$$

Since $\bar{y}$ is unbiased for $\bar{Y}$, the sample proportion p is unbiased for P. From (2.12) and (4.4),

$$V(p) = \frac{N-n}{N-1}\frac{PQ}{n}. \tag{4.7}$$

From (2.14) and (4.6), the estimator of this variance is

$$v(p) = \frac{N-n}{N}\frac{pq}{n-1}. \tag{4.8}$$

The actual and estimate of the standard errors of p are obtained from the square roots of (4.7) and (4.8), respectively.

An unbiased estimator for Q is given by the sample proportion q. Its variance and estimator of variance are the same as (4.7) and (4.8), respectively. It is directly shown in Appendix A4 that the sample proportion and the variance in (4.8) are unbiased.

The probability of observing c and $(n-c)$ units in the sample from the C and $(N-C)$ units of the two groups is given by the **hypergeometric** distribution:

$$\Pr(c, n-c) = \binom{C}{c}\binom{N-C}{n-c} \Big/ \binom{N}{n}. \tag{4.9}$$

It is shown in Appendix A4 that for large N, (4.9) can be approximated by the **binomial** distribution:

$$\Pr(c) = \frac{n!}{c!(n-c)!} P^c Q^{n-c}. \tag{4.10}$$

For this distribution, $p = c/n$ is unbiased for P with $V(p) = PQ/n$, and an unbiased estimator of this variance is $v(p) = pq/n$. Note that for large N, (4.7) becomes the same as PQ/n. If n is also large, the difference of (4.8) from pq/n will be negligible.

Example 4.1. Hospital ownership: The American Hospital Association *Guide to Health Care Field* (1988) contains information on the hospitals regarding the number of physicians, occupancy rates, annual expenses, and related characteristics. In short-term hospitals, patients' stay is limited to 30 days. For medical and surgical services, New York State has 217 short-term hospitals. Of these hospitals, 150 are owned by individuals, partnerships, or churches. The remaining 67 are operated by federal, state, local, or city governments. In a random sample of 40 selected from the 217 hospitals, 29 were found to be in the first group and 11 in the second.

The estimate of the proportion of the hospitals in the first group is $p = 29/40 = 0.725$, or close to 73%. From (4.8),

$$v(p) = \frac{(217-40)}{217}\frac{(0.725)(0.275)}{39} = 0.0042$$

and S.E.(p) = 0.0648. From these figures, an estimate of the proportion of the hospitals in the second group is 0.275, and it has the same S.E.

4.3 Total number or count

An unbiased estimator of the total number is $\hat{C} = Np$. Its variance and estimator of variance are given by

$$V(\hat{C}) = N^2 V(p) = \frac{N^2(N-n)}{N-1}\frac{PQ}{n} \tag{4.11}$$

and

$$v(\hat{C}) = N^2 v(p) = N(N-n)\frac{pq}{n-1}, \tag{4.12}$$

respectively.

From the sample in Example 4.1, an estimate for the total number of hospitals in the first group is $\hat{C} = 217(0.725) = 157$, and from (4.12), S.E.($\hat{C}$) = 217(0.0648) = 14.

4.4 Confidence limits

If N and n are large, $(p - P)/\sqrt{V(p)}$ approximately follows the standard normal distribution. For large N, this approximation becomes increasingly valid as n is larger than 25 or 30 and P is close to 0.5.

With this approximation, the lower and upper limits for P are obtained from

$$p \pm Z.\ \text{S.E.}(p), \tag{4.13}$$

where S.E.(p) is obtained from (4.8) and Z as before is the $(1 - \alpha)$ percentile of the standard normal distribution. Since a discrete variable is being approximated by a continuous one, the *continuity correction* $1/2n$ can be added to the upper limit in (4.13) and subtracted from the lower limit. If n is large, the effect of this correction becomes negligible.

Example 4.2. Hospital ownership: For the illustration in Example 4.1, S.E.(p) = 0.0648. Approximate 95% confidence limits for the proportion in the first group are given by 0.725 ± 1.96(0.0648); that is, the lower and upper limits are 0.598 and 0.852, respectively. The limits for the total number of hospitals in this group are 217(0.598) = 130 and 217(0.852) = 185.

4.5 Sample sizes for specified requirements

Similar to the procedures in Section 2.12 for the mean and total of a quantitative variable, sample sizes to estimate P or C can be found with specified requirements. The following types of specifications are usually required for estimating proportions in industrial research as well as for estimating percentages in political and public polls.

1. The variance of p should not exceed V; that is, S.E.(p) should not exceed $\sqrt{V}$. With this specification, from the expression for the variance in (4.7), the minimum sample size required is given by

$$n = \frac{n_l}{1 + (n_l - 1)/N}, \tag{4.14}$$

where $n_l = PQ/V$ is the size needed for a large population.

For each of the following specifications, the sample size n_l for large N is presented, and the required sample size is obtained from (4.14).

2. The error of estimation $|p - P|$ should not exceed e, except for an α% chance. This specification is the same as

$$\Pr[|p - P| / \text{S.E.}(p) \leq e / \text{S.E.}(p)] = 1 - \alpha. \tag{4.15}$$

Assuming that $|p - P|/\text{S.E.}(p)$ follows the standard normal distribution, the right-hand side term inside the square brackets is the same as the standard normal deviate Z corresponding to the $(1 - \alpha)$ probability. Hence

$$n_l = Z^2 PQ / e^2. \tag{4.16}$$

3. Confidence width for P, for a confidence probability $(1 - \alpha)$, should not exceed w. With the normal approximation, the width of the

interval in (4.13) is $2Z$ S.E.(p). Since it should not exceed w,

$$n_1 = 4Z^2PQ/w^2, \tag{4.17}$$

and n is obtained from (4.14).

4. The coefficient of variation of p, $\sqrt{V(p)}/P$, should not exceed C. In this case,

$$n_1 = Q/PC^2. \tag{4.18}$$

5. The relative error $|p - P|/P$ should not exceed **a**, except for an α% chance. For this requirement,

$$n_1 = Z^2Q/P\mathbf{a}^2. \tag{4.19}$$

The first three specifications require prior knowledge of the unknown proportion. In contrast, the last two specifications can be made without such information. The solutions to the sample sizes for all these five specifications, however, depend on the unknown proportions. If such information is not available, the sample sizes for the first three requirements can be found with $P = Q = 0.5$, since $V(p)$ takes its maximum in this case. For the remaining two specifications, relatively larger sample sizes are needed when P is between 0 and 0.5.

To estimate the percentages of qualitative characteristics, the proportion p, its S.E., and the specifications are expressed in percentages. The same solutions for the sample sizes are obtained again.

The first specification is equivalent to the requirement that the variance of $\hat{C}$ should not exceed N^2V. Thus, n in (4.14) is also the sample size needed for estimating C with this requirement. Similarly, the sample sizes with the remaining four specifications satisfy the equivalent requirements for the estimation of C.

As in the case of the population mean and total, larger sample sizes are needed to estimate the proportions, percentages, and total numbers with smaller S.E. values, errors of estimation, confidence widths, or relative errors, and with larger confidence probabilities.

Example 4.3. Sample size: To estimate the proportion P of a large population having a given attribute, with an error not exceeding 5%

except for a 1 in 20 chance, from (4.16),

$$n_l = (1.96)^2 PQ/(0.05)^2.$$

If it is known that P is not too far from 0.5, $n_l = 384$ from this solution.

If $P = 0.5$ and $\alpha = 0.05$ as above, and the relative error should not exceed 1%, n_l from (4.19) is again equal to 384.

With the above justifications, samples of size around 400 are frequently used in several political, public, and marketing surveys, even when the population size is very large. Estimates of the proportions with high precisions will be obtained, if the population is first divided into homogeneous groups or strata and sample sizes for each group are determined through the above types of requirements. With such a procedure, estimates for a population proportion are obtained by suitably weighting the estimates from the different groups.

4.6 Political and public polls

A number of polls are conducted during election times to summarize the preferences of the voters and to predict the winners. The results of five of the 1992 presidential election polls are presented in Table 4.1. These pre-election polls were conducted from October 28 to November 1 by the Gallup and Harris organizations, major newspapers, and television networks. Sample sizes for these surveys ranged from 982 to 1975. The last entry in the table is from an exit poll at 300 polling

Table 4.1. Presidential election—Percentages of voters for the three candidates.

Description	Clinton	Bush	Perot
	Pre-election polls		
n = 9,115	44	37	16
1,562	44	36	14
1,975	44	38	17
2,248	44	35	15
982	44	36	15
	Election day		
n = 15,490	43	38	19

Source: The New York Times, November 4 and 5, 1993.

Table 4.2. Public surveys—Percentages of responses for different items.

Are you a better than average driver? $n = 597$	Better	Worse	Same
	49	3	40
Problem with today's movies $n = 597$	Sex	Violence	Both
	13	44	38
President Clinton's health-care program $n = 500$	Favor	Oppose	Not Sure
Sept. 23	57	31	12
Oct. 28	43	36	21

Source: Time Magazine of December 1992 for the first two surveys and of November, 1993 for the last two.

Table 4.3. Classification of voters' participation in the primaries on Super Tuesday, March 7, 2000; (percentages).

			No. of Years of College			
	Men	Women	<4	4	>4	None
Democrats	42	58	21	22	31	26
Republicans	55	45	26	24	20	30

Source: Voters News Service poll of 1698 Democrats and 1432 Republicans at the primary election places throughout New York State on Super Tuesday, March 7, 2000; reported in *New York Times* of March 9, 2000, page A20.

stations on a total of 15,490 voters, conducted on election day November 3, 1992 by Voters Research and Surveys.

Results of four surveys of general interest to the public conducted by Yankelovich Inc. are presented in Table 4.2. The first two surveys were conducted in December 1992, and the last two in September and October of 1993.

Classification of 1698 Democratic and 1432 Republican registered voters who participated in the primaries on Super Tuesday, March 7, 2000, is presented in Table 4.3.

For these surveys, the error of estimation $\mathbf{e} = |p - P|$, which is also referred to as the **sampling error** or **margin of error**, is computed for a confidence probability of 95% with the normal approximation; that is, as described in Section 4.5 for the second criterion, **e** is obtained from equating **e**/S.E.(p) to 1.96. The data from these tables will be analyzed in the following sections and in the exercises.

4.7 Estimation for more than two classes

In Tables 4.1 and 4.2, responses to the surveys are classified into three groups, and they do not add to 100% for some of the surveys. The remaining percentages of the persons selected in the samples probably refused to respond to the surveys or they were not sure of their responses. If the nonresponse group is included, the populations for these surveys can be considered to consist of four classes.

As presented in Table 4.2, in the sample of 500 persons contacted on September 23, the health-care program was favored by 57% and opposed by 31%. The remaining 12% were not sure of their preferences.

In Table 4.3, in addition to the Men–Women classification, voters are classified into four groups according to the number of years of college education.

Probabilities of the outcomes

To represent the above type of situations, let $C_1, C_2, \ldots, C_k$ denote the number of units of a population belonging to k classes defined according to a qualitative characteristic. The corresponding proportions of the units are $P_1 = C_1/N$, $P_2 = C_2/N, \ldots, P_k = C_k/N$. The probability of observing $c_1, c_2, \ldots, c_k$ units in the classes in a random sample of size n is

$$P(c_1, c_2, \ldots, c_k) = \binom{C_1}{c_1}\binom{C_2}{c_2} \cdots \binom{C_k}{c_k} \Big/ \binom{N}{n}. \tag{4.20}$$

This distribution is an extension of (4.9) for $k \geq 3$ classes.

The estimates of the proportions and their variances can be obtained through the procedures in Section 4.2. The sample proportion $p_i = c_i/n$, $i = 1, 2, \ldots, k$, is unbiased for P_i. Its variance $V(p_i)$ is obtained from (4.7) by replacing P with P_i. Similarly, the estimator of variance $v(p_i)$ is obtained from (4.8) by replacing p with p_i. Unbiased estimator for the number of units in the ith class is $\hat{C}_i = Np_i$. The S.E. of $\hat{C}_i$ is given by $N\sqrt{V(p_i)}$ and it is estimated from $N\sqrt{v(p_i)}$. As in the case of two classes, confidence limits for the proportions and the numbers of units in the different classes are obtained by adding and subtracting Z times the respective S.E. values to the estimates.

Multinomial approximation

If N is large, the probability in (4.20) becomes

$$P(c_1,c_2,\ldots,c_k) = \frac{n!}{c_1!c_2!\ldots c_k!}P_1^{c_1}P_2^{c_2}\ldots P_k^{c_k}, \tag{4.21}$$

which is the **multinomial** distribution. This expression refers to the chance of observing c_i units in the ith class in n independent trials, with the probabilities P_i remaining constant from trial to trial. For this distribution, $p_i = a_i/n$ is unbiased for P_i. Its variance and estimator of variance are $V(p_i) = P_i(1-P_i)/n$ and $v(p_i) = p_i(1 - p_i)/n$.

Example 4.4. Presidential candidates: The sample size for the fourth survey in Table 4.1 is relatively large. Observations from this survey can be used for estimating the standard errors of the sample proportions and to find confidence intervals for the proportions of the population of voters favoring each of the three candidates.

For Mr. Clinton, an estimate of the proportion is 0.44. Its sample variance is (0.44)(0.56)/2247 = 0.0001 and hence the S.E. of the sample proportion is 0.01. The 95% confidence limits for the population proportion of the voters favoring him are 0.44 − 1.96(0.01) = 0.42 and 0.44 + 1.96(0.01) = 0.46. Thus, an estimate for the percentage of the voters favoring Mr. Clinton is 44, with the 95% confidence interval of 42 to 46%.

Similarly, 35% of the voters are for Mr. Bush, and the confidence limits are (33, 37)%. For Mr. Perot, the estimate is 15% and the confidence limits are (14, 16)%.

4.8 Combining estimates from surveys

The presidential election is an example in which several surveys are conducted independently to estimate the same percentages and total numbers in different classes of a population. Consider a population proportion P. Let p_i, $i = 1,2,\ldots,k$, denote its estimators from k independent samples of sizes n_i. The variances $V(p_i)$ of these estimators and their sample estimates $v(p_i)$ are obtained from (4.7) and (4.8) by replacing n by n_i.

An unbiased estimator for P is given by the weighted average $p = \sum(n_i/n)p_i$, where $n = \sum n_i$ is the overall sample size. The variance of p is $\sum(n_i/n)^2V(p_i)$, which is smaller than the variances of the individual p_i, and it is estimated from $\sum(n_i/n)^2v(p_i)$. If the population size and the sample sizes are large, the variance of p can be estimated from $\sum(n_i/n)^2p_i(1 - p_i)/n_i$.

Example 4.5. Combining independent estimates: Estimates of the percentages for the three candidates can be obtained by combining the five surveys in Table 4.1 assuming that they are independent. The total sample size is $n = 15{,}882$.

For Mr. Clinton, since each p_i is equal to 0.44, the overall estimate is also $p = 0.44$, and $v(p) = 0.44(0.56)/15{,}882$ and S.E.$(p) = 0.0039$. The 95% confidence intervals for the proportion are $0.44 - 1.96(0.0039) = 0.43$ and $0.44 + 1.96(0.0039) = 0.45$. Thus, the overall estimate is 44% and the confidence limits are (43, 45)%.

Similarly, for Mr. Bush, the overall estimate is 37% and the confidence limits are (36, 38)%. For Mr. Perot, the overall estimate is 15% and the confidence limits are (14, 16)%. As can be seen for each of the three candidates, the standard errors of the combined estimates are smaller than the standard errors found in Example 4.4 from any one of the surveys, and the confidence widths are smaller.

4.9 Covariances of proportions

As will be seen below, for comparing the proportions in the classes, variances as well as covariances of their estimates will be needed. To obtain expressions for the covariances, note that

$$V(p_1 + p_2 + \cdots + p_k) = V(p_1) + V(p_2) + \cdots + V(p_k) + 2\mathrm{Cov}(p_1, p_2) + 2\mathrm{Cov}(p_1, p_3) + \cdots. \tag{4.22}$$

Since $(p_1 + p_2 + \cdots + p_k) = 1$, the variance on the left-hand side vanishes. The variances $V(p_i)$ are given by $(N - n)P_i(1 - P_i)/(N - 1)n$. Substituting these expressions in (4.22), for $(i \neq j)$, $(i, j) = 1, 2, \ldots, k$,

$$\mathrm{Cov}(p_i, p_j) = -\frac{N - n}{N - 1}\frac{P_i P_j}{n} \tag{4.23}$$

Similarly, the estimator of this covariance is

$$\hat{\mathrm{Cov}}(p_i, p_j) = -\frac{N - n}{N}\frac{p_i p_j}{n - 1} \tag{4.24}$$

The expressions in (4.23) and (4.24) can also be obtained directly from the probability distribution in (4.20).

4.10 Difference between proportions

Comparisons of the proportions and the numbers of units in the different classes are frequently of interest. An unbiased estimator of $(P_i - P_j)$ is $(p_i - p_j)$. Its variance and estimator of variance are

$$\begin{aligned} V(p_i - p_j) &= V(p_i) + V(p_j) - 2\mathrm{Cov}(p_i, p_j) \\ &= \frac{N-n}{(N-1)n}[P_i(1-P_i) + P_j(1-P_j) + 2P_iP_j] \end{aligned} \quad (4.25)$$

and

$$v(p_i - p_j) = \frac{N-n}{N(n-1)}[p_i(1-p_i) + p_j(1-p_j) + 2p_ip_j]. \quad (4.26)$$

An estimate for the difference $(C_i - C_j)$ is $N(p_i - p_j)$. The sample S.E. of this estimate is obtained from $N\sqrt{v(p_i - p_j)}$. As can be seen from these expressions, the standard errors of the estimates of $(P_i - P_j)$ and $(C_i - C_j)$ are larger than the standard errors of the individual proportions or totals.

Example 4.6. Educational standards: As reported in Newsweek magazine (May, 1988), in a survey conducted by the Gallop organization in 1987, in a sample of 542 college students, 62% thought that they were receiving a "better" education than their parents' generation, 7% said "worse," and 25% thought that it was about the same. The remaining 6% were the "Don't knows."

An estimate of the difference of the proportions in the first and third categories is $(0.62 - 0.25) = 0.37$ or 37%. With the assumption of a simple random sample for the above survey and a large size for the population, from (4.26), the variance of this estimate is $[0.62(0.38) + 0.25(0.75) + 2(0.62)(0.25)]/541 = 13.55 \times 10^{-4}$, and the S.E. is 0.0368 or about 3.7%. The 95% confidence limits for the difference are given by $0.37 \pm 1.96(0.0368)$, that is, 30 and 44%.

4.11 Sample sizes for more than two classes

When there are more than two classes, the sample size required to estimate the proportions and totals can be found from the type of specifications described in Section 4.5. Similarly, sample size to estimate the difference $(P_i - P_j)$ of two proportions can be found by replacing $V(p_i)$ by $V(P_i - P_j)$ in the corresponding expressions. If no information

on the proportions is available, as before the maximum sample size is obtained by replacing each proportion by 0.5.

Example 4.7. Sample sizes for two or more classes: Consider more than two classes, as in the case of the political and public polls described in Section 4.6, and the problem of estimating their proportions. If the error of estimation for estimating any of the proportions should not exceed 0.05 except for a 5% chance, the sample size is obtained from (4.16). If it is known, for example, that the maximum of the percentages is 45, the required sample size is given by $(1.96)^2(0.45)(0.55)/(0.05)^2 = 380.32$ or 381 approximately. If no information is available on the percentages, as in Example 4.3, a sample size of at least 384 would be needed.

To estimate the difference $(P_i - P_j)$ of two proportions with an error not exceeding 0.05 except for a 5% chance, the required sample size is obtained by equating $\mathbf{e}/\text{S.E.}(p_i - p_j)$ to 1.96. With 0.45 for each proportion, from (4.25), the variance of $(p_i - p_j)$ for large N equals $[(0.45)(0.55) + (0.45)(0.55) + 2(0.45)(0.55)] = 0.9/n$. Thus, the required sample size is equal to $(1.96)^2(0.9)/(0.05)^2 = 1383$. With 0.5 for each proportion, the sample size would be $(1.96)^2/(0.05)^2 = 1537$ approximately. In either case, sample sizes required for estimating the difference are larger than the sizes for the individual proportions. This result is expected since the S.E. of the difference of the sample proportions is larger than the S.E. of an individual sample proportion.

4.12 Estimating population size

Throughout the previous sections, the population size N was assumed to be known. To estimate the sizes of animal, fish, and other types of moving populations, the **capture–recapture** method can be used. In this procedure, a sample of n_1 units observed from the population is tagged or marked and released. The number of units c appearing again in an independent sample of n_2 units is noted. The probability distribution of these common units is

$$P(c) = \binom{n_1}{c}\binom{N-n_1}{n_2-c}\Big/\binom{N}{n_2} = \binom{n_2}{c}\binom{N-n_2}{n_1-c}\Big/\binom{N}{n_1}, \qquad (4.27)$$

where $0 \le c \le \min(n_1, n_2, N - n_1 - n_2)$.

The proportion of the units tagged in the first sample is n_1/N. An unbiased estimator of this proportion is c/n_2. Hence, $\hat{N} = n_1 n_2/c$ can be considered to be an estimator of N. Another motivation for $\hat{N}$ arises

from noting that $E(c) = n_1 n_2/N$. It is, however, an overestimate of N since $E(1/c) \geq 1/E(c)$.

For large N, (4.27) approximately becomes the binomial distribution

$$P(c) = \binom{n_2}{c} P^c Q^{n_2 - c} \tag{4.28}$$

where $P = (n_1/N)$. Now, (c/n_2) is an unbiased estimator of this proportion. From this result, again $n_1 n_2/c$ can be considered an estimator for N.

Exercises

4.1. Enrollments for higher education beyond high school for the 50 states in the U.S. and the District of Columbia (DC) are available in the *Statistical Abstracts of the United States*; figures for 1990 are available in the 1992 publication. In a random sample of 10 from 49 states and DC, excluding the two largest states, New York and California, enrollments in 1990 exceeding 100,000 were observed for public institutions in six states and for private institutions in one state. (a) For each type of institution, find the estimate and its S.E. for the number of states in which enrollments exceeded 100,000. (b) Estimate the difference of the above numbers for the two types and find the S.E. of the estimate.

4.2. If no information on P is available, find the error of estimation **e** in (4.15) for (a) $n = 1000$ and (b) $n = 2000$. Consider $\alpha = 0.05$ for both cases.

4.3. The sampling errors for the first two and last two public polls in Table 4.2 were reported to be 4 and 4.5%, respectively. Explain how these results are obtained, noting that the estimate in each case should not differ from the population percentage by the specified amount of error except for a 5% chance.

4.4. From the sample data in Table 4.2, for the change from September to October in the percentage favoring the health care program, find (a) the estimate, (b) its S.E., and (c) the 95% confidence limits.

4.5. From the results in Table 4.2 for September, for the percentage not in favor, that is, opposing and not sure, find (a) the estimate, (b) its S.E., and (c) the 95% confidence limits. (d) Test the hypothesis at $\alpha = 0.05$ that the percentage favoring is not more than 10% from the percentage not favoring.

4.6. From the figures in Table 4.3, for the participation of both the Democrats and Republicans together, find the estimate, its S.E., and the 95% confidence limits for the (a) percentage of men, (b) percentage of women, and (c) difference in the percentages for men and women.

4.7. From the figures in Table 4.3, for the percentage of the participants in the primaries having at least some college education, find the estimate, its S.E., and the 95% confidence limits for (a) Democrats, (b) Republicans, and (c) both Democrats and Republicans together. (d) Describe how one can find these estimates, standard errors, and confidence limits from the percentages of those who do not have any college education.

4.8. For the survey on 1038 teenagers described in Section 1.8, 75 and 81% of the boys and girls, respectively, emphasized the importance of good grades. For 53 and 41% of the boys and girls, it was important to have a lot of friends. (a) For both good grades and a lot of friends, find the 95% confidence limits for the difference in the percentages for boys and girls. (b) For good grades and for having a lot of friends, test the hypothesis at the 5% level that there is no significant difference between boys and girls. Assume that the responses were obtained from equal number, 519, of boys and girls.

4.9. One university has 1000 students in each of the four classes. The percentages of the Freshman–Sophomore, Junior, and Senior classes expressing interest in professional training after graduation were guessed to be 20, 50, and 80%, respectively. (a) For each of these three groups, find the sample sizes required to estimate the percentage if the estimate should not differ from the actual value by not more than 20% of the actual value except for a 5% chance, and present the reason for the differences in the

sample sizes. (b) Find the sample sizes needed for each of the three groups for estimating the above percentage if the error of estimation should not exceed 10% except for a 5% chance, and present the reason for the differences in the sample sizes.

4.10. Among the 637 families in a village in New York State, 241 have children currently attending school. In the remaining 396 families, either the children have completed schooling or they have no children. In a straw poll, 14 families from the first group favored the proposal to increase the school budget, two had no opinion, but nine opposed. The corresponding figures for the second group are 9, 3, and 12. For the proportion and number of families in the village who have no opinion, find (a) the estimate, (b) its S.E., and (c) the 95% confidence limits. For the difference in the proportions and numbers of the families in the two groups favoring the proposal, find (a) the estimate, (b) its S.E., and (c) the 95% confidence limits.

4.11. With the preliminary estimates obtained from the data in Exercise 4.10, find the sample sizes needed to estimate the percentage of the families favoring the proposal (a) with an error less than 10% except for a 1 in 20 chance, (b) with the error not exceeding 1/10 of the actual percentage except for a 5% chance, and (c) with the confidence width for a 90% confidence probability not exceeding 10%.

4.12. Using the preliminary estimates from the data in Exercise 4.10, find the sample sizes required to estimate the total number of families favoring the proposal (a) if the error of estimation should not exceed 10% of the actual number except for a 5% chance and (b) if the coefficient of variation of the estimate should not be more than 5%.

4.13. At the end of a cultural conference at a university, 110 faculty, 250 students, and 200 other participants returned the questionnaires on their opinions regarding the conference. Among these three groups, 80, 120, and 100 were in favor of continuing the conference every year. Denote the actual proportions in the three groups favoring the continuation by P_1, P_2, and P_3. Using the preliminary estimates

from these responses, find the overall sample sizes needed (a) to estimate $(P_1 - P_2)$ with an S.E. not exceeding 7%, and (b) to estimate the difference between the proportions for any two groups with an S.E. not exceeding 7%. The university community consists of 500, 6000, and 2000 persons in the above three categories.

4.14. In the survey described in Exercise 3.9, 12 of the 15 respondents suggested that the campus should have more tennis courts. Thus, for this characteristic, the sample estimate of the proportion for the respondents is $p_1 = 12/15 = 0.8$. For the population proportion, three estimates may be considered: (1) p_1, (2) substitute p_1 for each of the nonrespondents, and (3) substitute zero for each of the nonrespondents. Find expressions for the biases, variances, and MSEs of these three procedures.

4.15. From (4.20), show that $E(c_1c_2) = n(n-1)C_1C_2/N(N-1)$. Using this result, (a) derive the covariance in (4.23) and (b) show that the expression in (4.24) is unbiased for this covariance.

Appendix A4

Unbiasedness of the sample proportion and variance

With the probability in (4.9),

$$E(c) = \sum_c c \Pr(c, n-c) = nP$$

and hence $P = c/n$ is unbiased for P.

Similarly,

$$E[c(c-1)] = \frac{n(n-1)}{N(N-1)} C(C-1),$$

which can be expressed as

$$E[np(np-1)] = \frac{n(n-1)}{N(N-1)} NP(NP-1),$$

From this expression,

$$E(p^2) = \left[\frac{N-n}{N-1} + \frac{N(n-1)P}{N-1}\right]\frac{P}{n}.$$

Finally writing $V(p)$ as $E(p^2) - P^2$, the expression for the variance becomes the same as (4.7).

From this result, an unbiased estimator for P^2 is given by $[(N-1)np^2 - (N-n)p]/N(n-1)$.

Since $PQ = P - P^2$, its unbiased estimator from the above results is given by $p - [(N-1)np^2 - (N-n)p]/N(n-1)$, that is,

$$\widehat{PQ} = \frac{(N-1)n}{N(n-1)}pq.$$

By utilizing this result, (4.8) becomes the unbiased estimator of the variance in (4.7).

Binomial approximation

The hypergeometric distribution in (4.9) can be expressed as

$$\Pr(c, n-c) = \frac{n!}{c!(n-c)!}\frac{C!(N-C)!(N-n)!}{(C-c)!(N-C-n+c)!N!}$$

If N is large, this expression becomes

$$\Pr(c, n-c) = \frac{n!}{c!(n-c)!}P^cQ^{n-c},$$

which is the **binomial** distribution.

This is the probability for c *successes* in n independent trials, with the probability of success P remaining the same at every trial—the **Bernoulli** trials. For this distribution, $p = c/n$ is unbiased for P. Its variance and estimator of variance are $V(p) = PQ/n$ and $v(p) = pq/n$.