

PART I

Basic Sampling

Simple Random Sampling

Simple random sampling, or *random sampling without replacement*, is a sampling design in which n distinct units are selected from the N units in the population in such a way that every possible combination of n units is equally likely to be the sample selected. The sample may be obtained through n selections in which at each step every unit of the population not already selected has equal chance of selection. Equivalently, one may make a sequence of independent selections from the whole population, each unit having equal probability of selection at each step, discarding repeat selections and continuing until n distinct units are obtained.

A simple random sample of $n = 40$ units from a population of $N = 400$ units is depicted in Figure 2.1. Another simple random sample, just as likely as the first to be selected, is shown in Figure 2.2. Each such combination of 40 units has equal probability of being the sample selected. With simple random sampling, the probability that the i th unit of the population is included in the sample is $\pi_i = n/N$, so that the inclusion probability is the same for each unit. Designs other than simple random sampling may give each unit equal probability of being included in the sample, but only with simple random sampling does each possible *sample* of n units have the same probability.

2.1. SELECTING A SIMPLE RANDOM SAMPLE

A simple random sample may be selected by writing the numbers 1 through N on N pieces of paper, putting the pieces of paper in a hat, stirring them thoroughly, and, without looking, selecting n of the pieces of paper without replacing any. The sample consists of the set of population units whose labels correspond to the numbers selected. To reduce the labor of the selection process and to avoid such problems as pieces of paper sticking together, the selection is more commonly made using a random number table or a computer “random number” generator.

To select a simple random sample of n units from the N in the population using a random number table, one may read down columns of digits in the table

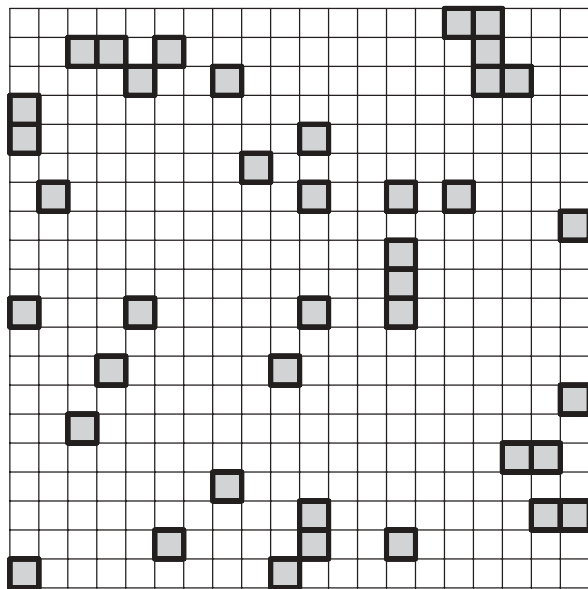


Figure 2.1. Simple random sample of 40 units from a population of 400 units.

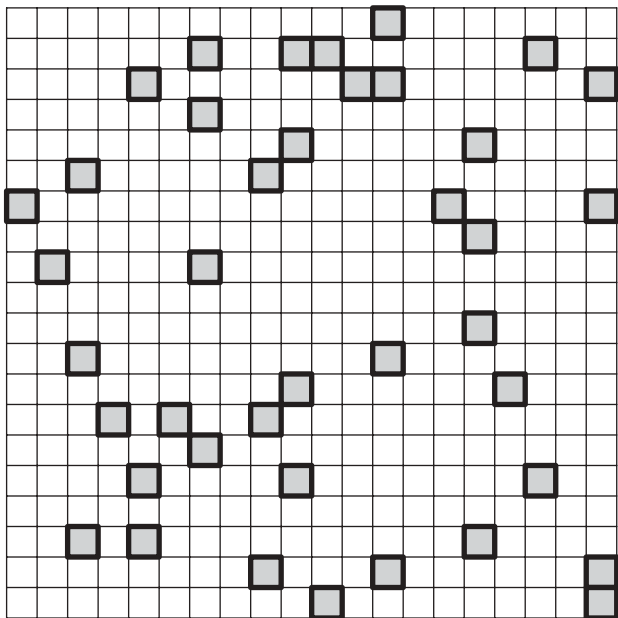


Figure 2.2. Another simple random sample of 40 units.

starting from a haphazard location. As many columns of the table as $N - 1$ has digits are used. When using three columns, the digits “000” would be interpreted as unit 1000. When using the table, repeat selections and numbers greater than N are ignored, and selection continues until n distinct units are obtained.

The basic random number generator on most computers produces decimal fractions uniformly distributed between zero and 1. The first few digits after the decimal place in such numbers can be used to represent unit label numbers.

Table 2.1 lists 285 uniform random numbers, each shown to 10 digits after the decimal point, produced by a computer random number generator. Suppose that we wish to select a simple random sample of $n = 10$ units from a population of $N = 67$ units. Starting from the first entry in Table 2.1 and reading pairs of digits downward (it would also be valid to read across rows of the table, to use pairs of digits other than the first after the decimal point, or to start at a point other than the beginning of the list), the first pair is 99. Since there is no unit 99 in the population, this entry is passed over, and the first unit selected to be in the sample is unit 21. Continuing down the column, the sample selected consists of the units 21, 12, 1, 15, 29, 43, 30, 63, 2, and 8. In making the selection, note that entries 68, 76, 86, 97, and 100 (represented by the pair 00) were passed over since each is larger than N . Entry 43 was passed over the second time it appeared, so the sample contains 10 distinct units. Many computer systems include facilities for direct random selection without replacement of n integers between 1 and N , eliminating the tedium of passing over repeat values or values larger than N . For example, in either of the statistical systems R or S-PLUS, the command “`s <- sample(1:27502, 12500, replace = F)`” selects $n = 12,500$ integers at random without replacement from the set of integers from 1 to 27,502 and stores the selected numbers in “s.”

2.2. ESTIMATING THE POPULATION MEAN

With simple random sampling, the sample mean \bar{y} is an unbiased estimator of the population mean μ . The population mean μ is the average of the y -values in the whole population:

$$\mu = \frac{1}{N}(y_1 + y_2 + \cdots + y_N) = \frac{1}{N} \sum_{i=1}^N y_i \quad (2.1)$$

The sample mean \bar{y} is the average of the y -values in the sample:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.2)$$

Also with simple random sampling, the sample variance s^2 is an unbiased estimator of the *finite-population variance* σ^2 . The finite-population variance is defined as

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2 \quad (2.3)$$

Table 2.1: Uniform Random Numbers

.9915338159	.3376058340	.1529208720	.0008221702	.3645994067
.2110764831	.4482254982	.0259101614	.1159885451	.5011445284
.1215928346	.4434396327	.1677099317	.5284986496	.9135305882
.0125039294	.2536827028	.1724499613	.5171836615	.5422329903
.1583184451	.4694896638	.9516881704	.3874872923	.0451180041
.2974444926	.9606751800	.2988916636	.7681296468	.3288438320
.4321415126	.9025109410	.6112304330	.4916386008	.8434410095
.3065150678	.5485164523	.6078377366	.1443793625	.7657701969
.6806892753	.0791656822	.7079550028	.7407252192	.7297828197
.7614942193	.4598654807	.8545978069	.4847860932	.7846541405
.8696339726	.2160511613	.5071278811	.0302107912	.3910638690
.4398060441	.0101473443	.0496022329	.2955447733	.6359770298
.9754472375	.0900140777	.9543433189	.7030580044	.6982350349
.6345051527	.9645981193	.4215144813	.8500274420	.4303097129
.0047403700	.9751796722	.6224800944	.4581535459	.3851253986
.0205896683	.2392801940	.0118337637	.6197799444	.9798330665
.0894387960	.1349214613	.0790547207	.1108237952	.1181035042
.6207187772	.4988264143	.9772401452	.2934628427	.7792176604
.8887537122	.3153925836	.4549961388	.3680104315	.8818087578
.3764214814	.6713073850	.9082747102	.3485270441	.7828890681
.9147837162	.4565998316	.2507463396	.8603917360	.3503700197
.7551217675	.6151723266	.6706758142	.9292267561	.7541347742
.4477638602	.4369836152	.4551322758	.8340566158	.6796288490
.8799548149	.5218108892	.2309677154	.6433401108	.0874217674
.6529608965	.9821792245	.8369561434	.8693770766	.3227941990
.9485814571	.7658874393	.5788805485	.8377626538	.1910941452
.9316777587	.5495033860	.7132855058	.9236876369	.1685705334
.6445560455	.1993282586	.1627506465	.0411975421	.0192697253
.0773160681	.6400896907	.4214436412	.5431558490	.5692960024
.8540129066	.5267632008	.6384039521	.4066059291	.0482674502
.6418970227	.2250400186	.6437576413	.2099322975	.3629093170
.1715016663	.1052204221	.6630748510	.1328498721	.1639286429
.1240955144	.0937742889	.4384917915	.4143532813	.8565336466
.9962730408	.1046832651	.1845341027	.7540032864	.8298202157
.1585547477	.7293077707	.7993465066	.7446641326	.5463740826
.7089923620	.1290157437	.8575667739	.0251938123	.7664318085
.3898053765	.9139558077	.3378374279	.2337769121	.4814206958
.7222445011	.6537817717	.1274980158	.0039445930	.3522033393
.1698853821	.5726385117	.7305127382	.2965210974	.2888952196
.5344746709	.2255166918	.0169686452	.5906063914	.9546776414
.7548384070	.0843338221	.8771440983	.7653347254	.5916480422
.1039589792	.6858401299	.6389055848	.9076186419	.8857548237
.5081589222	.2550631166	.1969931573	.0558514856	.6456795335
.3169104457	.5660375357	.6318614483	.1304887086	.4802035689
.6693667173	.9299270511	.8694118261	.2035958767	.9613003135

Table 2.1: Uniform Random Numbers (Continued)

.3214286268	.8198484778	.8971202970	.0275031179	.1577183455
.1545569003	.2482915521	.7872648835	.4376204610	.2435218245
.5372928381	.5366832614	.4940558970	.5881735682	.5513799191
.0131097753	.9373838305	.9739696383	.5421801805	.3240519464
.3482980430	.7070090175	.6941514015	.1654081792	.3356401920
.4537515640	.8378376961	.3140848875	.5731232762	.2575304508
.3538932502	.5364976525	.0633419156	.2484393269	.7877063751
.6873268485	.3285647929	.7112956643	.5748419762	.8346126676
.1625820547	.6026779413	.9953029752	.7957111597	.2106933594
.9141720533	.6276242733	.7062586546	.0587451383	.3998769820
.4099894762	.7787652612	.3133662939	.8499189615	.0682335123
.4036674798	.4339759648	.7664646506	.0310811996	.7275006175

The sample variance is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2.4)$$

The variance of the estimator \bar{y} with simple random sampling is

$$\text{var}(\bar{y}) = \left(\frac{N-n}{N} \right) \frac{\sigma^2}{n} \quad (2.5)$$

An unbiased estimator of this variance is

$$\widehat{\text{var}}(\bar{y}) = \left(\frac{N-n}{N} \right) \frac{s^2}{n} \quad (2.6)$$

The square root of the variance of the estimator is its standard error; the estimated standard error is in general not an unbiased estimator of the actual standard error.

The quantity $(N-n)/N$, which may alternatively be written $1 - (n/N)$, is termed the *finite-population correction factor*. If the population is large relative to the sample size, so that the sampling fraction n/N is small, the finite-population correction factor will be close to 1, and the variance of the sample mean \bar{y} will be approximately equal to σ^2/n . Omitting the finite-population correction factor in estimating the variance of \bar{y} in such a situation will tend to give a slight overestimate of the true variance. In sampling small populations, however, the finite-population correction factor may have an appreciable effect in reducing the variance of the estimator, and it is important to include it in the estimate of that variance. Note that as sample size n approaches the population size N in simple random sampling, the finite-population correction factor approaches zero, so that the variance of the estimator \bar{y} approaches zero.

2.3. ESTIMATING THE POPULATION TOTAL

To estimate the population total τ , where

$$\tau = \sum_{i=1}^N y_i = N\mu \quad (2.7)$$

the sample mean is multiplied by N . An unbiased estimator of the population total is

$$\hat{\tau} = N\bar{y} = \frac{N}{n} \sum_{i=1}^n y_i \quad (2.8)$$

Since the estimator $\hat{\tau}$ is N times the estimator \bar{y} , the variance of $\hat{\tau}$ is N^2 times the variance of \bar{y} . Thus,

$$\text{var}(\hat{\tau}) = N^2 \text{var}(\bar{y}) = N(N - n) \frac{\sigma^2}{n} \quad (2.9)$$

An unbiased estimator of this variance is

$$\widehat{\text{var}}(\hat{\tau}) = N^2 \widehat{\text{var}}(\bar{y}) = N(N - n) \frac{s^2}{n} \quad (2.10)$$

Example 1: Estimates from Survey Data. In an experimental survey of caribou on the Arctic Coastal Plain of Alaska, caribou were counted from an aircraft flying over selected lines across the study region (Davis et al. 1979; Valkenburg 1990). All caribou within 1/2 mile to either side of each line flown were recorded, so that each unit was a 1-mile-wide strip. A simple random sample of 15 north-south strips was selected from the 286-mile-wide study region, so that $n = 15$ and $N = 286$. The numbers of caribou in the 15 sample units were 1, 50, 21, 98, 2, 36, 4, 29, 7, 15, 86, 10, 21, 5, and 4.

The sample mean [using Equation (2.2)] is

$$\bar{y} = \frac{1 + 50 + \cdots + 4}{15} = 25.9333$$

The sample variance [using Equation (2.6)] is

$$s^2 = \frac{(1 - 25.93)^2 + (50 - 25.93)^2 + \cdots + (4 - 25.93)^2}{15 - 1} = 919.0667$$

The estimated variance of the sample mean [using Equation (2.6)] is

$$\widehat{\text{var}}(\bar{y}) = \left(\frac{286 - 15}{286} \right) \frac{919.07}{15} = 58.0576$$

so that the estimated standard error is $\sqrt{58.06} = 7.62$.

An estimate of the total number of caribou in the study region [using Equation (2.8)] is

$$\hat{\tau} = 286(25.9333) = 7417$$

The estimated variance associated with the estimate of the total [using Equation (2.10)] is

$$\widehat{\text{var}}(\hat{\tau}) = 286^2(58.0576) = 4,748,879$$

giving an estimated standard error of $\sqrt{4,748,879} = 2179$. □

2.4. SOME UNDERLYING IDEAS

The estimator \bar{y} is a random variable, the outcome of which depends on which sample is selected. With any given sample, the value of \bar{y} may be either higher or lower than the population mean μ . But the expected value of \bar{y} , taken over all possible samples, equals μ . Thus, the estimator \bar{y} is said to be *design-unbiased* for the population quantity μ , since the probability with respect to which the expectation is evaluated arises from the probabilities, due to the design, of selecting different samples.

Therefore, the unbiasedness of the sample mean of the population mean with simple random sampling does not depend on any assumptions about the population itself.

The variance estimates are similarly design-unbiased for their population counterparts. The actual variance of the estimator \bar{y} depends on the population through the population variance σ^2 . For a given population, however, a larger sample size n will always produce a lower variance for the estimators \bar{y} and $\hat{\tau}$.

Example 2: All Possible Samples. The ideas underlying simple random sampling can be illustrated with the sampling of a very small population. The object of the sampling is to estimate the number of persons attending a lecture. To make a very quick estimate, a random sample of $n = 2$ of the $N = 4$ seating sections in the lecture theater was selected, and the number of persons in each section selected was counted. The units (seating sections) were labeled 1, 2, 3, and 4, starting from the entrance.

Using random digits generated on a computer (four numbered pieces of paper in a hat would have done as well), the sample $\{1, 3\}$ was selected. There were 10 people in unit 1 and 13 people in unit 3. The data, which include the unit labels as well as the y -values in the sample, are $\{(i, y_i), i \in s\} = \{(1, 10), (3, 13)\}$.

The sample mean is $\bar{y} = (10 + 13)/2 = 11.5$. The estimate of the population total τ , the number of people attending the lecture, is $\hat{\tau} = N\bar{y} = 4(11.5) = 46$. The sample variance [using Equation (2.4)] is $s^2 = [(10 - 11.5)^2 + (13 - 11.5)^2]/(2 - 1) = 4.5$. The estimated variance of $\hat{\tau}$ [using Equation (2.10)] is $\widehat{\text{var}}(\hat{\tau}) = [(4)(4 - 2)(4.5)]/2 = 18$.

Had another sample of two units been selected, different values would have been obtained for each of these statistics. Since the population is so small, it is possible to look at every possible sample and the estimates obtained with each. Counting the number of people in the remaining seating sections, the population y -values were determined as summarized in the following table:

Unit, i	1	2	3	4
People, y_i	10	17	13	20

The population parameters are $\tau = 60$ people attending the lecture and $\mu = 15$ people per section on average; the finite-population variance is $\sigma^2 = 19.33$. With $N = 4$ and $n = 2$, there are $\binom{4}{2} = 6$ possible samples. Table 2.2 lists each of the possible samples s along the y -values y_s , and the estimates and the confidence interval (c.i.) obtained with each sample.

Because of the simple random sampling used, each possible sample has probability $P(s) = 1/6$ of being the one selected. An estimator such as $\hat{\tau}$ is a random variable whose value depends on the sample selected. The expected value of $\hat{\tau}$ with respect to the design is the sum, over all possible samples, of the value of the estimator for that sample times the probability of selecting that sample. Thus, the expected value of $\hat{\tau}$ is

$$E(\hat{\tau}) = 54\left(\frac{1}{6}\right) + 46\left(\frac{1}{6}\right) + 60\left(\frac{1}{6}\right) + 60\left(\frac{1}{6}\right) + 74\left(\frac{1}{6}\right) + 66\left(\frac{1}{6}\right) = 60$$

demonstrating for this population that the estimator $\hat{\tau}$ is indeed unbiased for the parameter τ under simple random sampling. Similarly, one can show directly for the other estimators that $E(\bar{y}) = \mu$, $E(s^2) = \sigma^2$, and $E[\widehat{\text{var}}(\hat{\tau})] = \text{var}(\hat{\tau})$. On the other hand, direct computation of the expected value, over all possible samples, of the sample standard deviation $s = \sqrt{s^2}$ gives $E(s) = 4.01$, while the population standard deviation is $\sigma = \sqrt{19.33} = 4.40$, so the sample standard deviation is not unbiased for the population standard deviation under simple random sampling.

The variance of $\hat{\tau}$ is the sum, over all possible samples, of the value of $(\hat{\tau} - \tau)^2$ times the probability of that sample. Thus, direct computation of the variance of $\hat{\tau}$

Table 2.2: Data for Example 2

Sample	y_s	\bar{y}	$\hat{\tau}$	s^2	$\widehat{\text{var}}(\hat{\tau})$
(1, 2)	(10, 17)	13.5	54	24.5	98
(1, 3)	(10, 13)	11.5	46	4.5	18
(1, 4)	(10, 20)	15.0	60	50.0	200
(2, 3)	(17, 13)	15.0	60	8.0	32
(2, 4)	(17, 20)	18.5	74	4.5	18
(3, 4)	(13, 20)	16.5	66	24.5	98

(using the data in Table 2.2) gives

$$\begin{aligned}\text{var}(\hat{\tau}) &= (54 - 60)^2 \left(\frac{1}{6}\right) + (46 - 60)^2 \left(\frac{1}{6}\right) + (60 - 60)^2 \left(\frac{1}{6}\right) \\ &\quad + (60 - 60)^2 \left(\frac{1}{6}\right) + (74 - 60)^2 \left(\frac{1}{6}\right) + (66 - 60)^2 \left(\frac{1}{6}\right) \\ &= 77.333\end{aligned}$$

□

2.5. RANDOM SAMPLING WITH REPLACEMENT

Imagine drawing n poker chips from a bowl of N numbered chips one at a time, returning each chip to the bowl before selecting the next. With such a procedure, any of the chips may be selected more than once. A sample of n units selected by such a procedure from a population of N units is called a *random sample with replacement*. The n selections are independent, and each unit in the population has the same probability of inclusion in the sample. Simple random sampling with replacement is characterized by the property that each possible *sequence* of n units—distinguishing order of selection and possibly including repeat selections—has equal probability under the design.

One practical advantage of sampling with replacement is that in some situations it is an important convenience not to have to determine whether any unit in the data is included more than once. However, for a given sample size n , simple random sampling with replacement is inherently less efficient than simple random sampling without replacement.

Let \bar{y}_n denote the sample mean of the n observations; that is,

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.11)$$

Note that if a unit is selected more than once, its y -value is utilized more than once in the estimator.

The variance of \bar{y}_n is

$$\text{var}(\bar{y}_n) = \frac{1}{nN} \sum_{i=1}^N (y_i - \mu)^2 = \frac{N-1}{nN} \sigma^2 \quad (2.12)$$

Thus, the variance of the sample mean with simple random sampling without replacement is lower, since it is $(N-n)/(N-1)$ times that of the sample mean of all the observations when the sampling is with replacement.

An unbiased estimate of the variance of \bar{y}_n is

$$\widehat{\text{var}}(\bar{y}_n) = \frac{s^2}{n} \quad (2.13)$$

The estimator \bar{y}_n depends on the number of times each unit is selected, so that two surveys observing exactly the same set of distinct units, but with different repeat selections, would in general yield different estimates. This situation can be avoided by using the sample mean of the distinct observations.

The number of distinct units contained in the sample, termed the *effective sample size*, is denoted v . Let \bar{y}_v be the sample mean of the distinct observations:

$$\bar{y}_v = \frac{1}{v} \sum_{i=1}^v y_i \quad (2.14)$$

The estimator \bar{y}_v is an unbiased estimator of the population mean. The variance of \bar{y}_v can be shown to be less than that of \bar{y}_n . However, it is still not as small as the variance of the sample mean under simple random sampling without replacement (see Cassel et al. 1977, p. 41). Even so, in some survey situations the practical convenience of sampling with replacement could allow a larger sample size to be used, resulting in improved precision for a given amount of time or expense.

Example 3: Random Sampling with Replacement. In a simple random sample with replacement with nominal sample size $n = 5$, the following y -values are obtained: 2, 4, 0, 4, 5. However, examination of the labels of the units in the sample reveals that one unit, the one with $y_i = 4$, was selected twice. The estimate of the population mean based on the sample mean of the five observations, not all of which are distinct [using Equation (2.11)], is

$$\bar{y}_n = \frac{2 + 4 + 0 + 4 + 5}{5} = 3.0$$

The estimate based only on the four distinct units in the sample [using Equation (2.14)] is

$$\bar{y}_v = \frac{2 + 4 + 0 + 5}{4} = 2.75$$

□

2.6. DERIVATIONS FOR RANDOM SAMPLING

Since the number of combinations of n distinct units from a population of size N is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \quad (2.15)$$

the design simple random sampling assigns probability $1/\binom{N}{n}$ to each possible sample s of n distinct units. The probability π_i that a given unit i is included in the sample is the same for every unit in the population and is given by $\pi_i = n/N$.

It is customary in sampling to write the y -values in the population as y_1, y_2, \dots, y_N and the y -values in the sample as y_1, y_2, \dots, y_n , and for most

purposes no confusion results from this simple notation. A more precise notation lists the y -values in sample s as $y_{s1}, y_{s2}, \dots, y_{sn}$, distinguishing, for example, that the first unit in the sample is not necessarily the same unit as the first unit in the population. With the more careful notation, the sample mean for sample s is written $\bar{y}_s = (1/n) \sum_{i=1}^n y_{si}$.

The expected value of the sample mean \bar{y} in simple random sampling is defined as $E(\bar{y}) = \sum \bar{y}_s P(s)$, where the summation is over all possible samples s of size n , and \bar{y}_s denotes the value of the sample mean for the sample s . This expectation may be computed directly, since $P(s) = 1/\binom{N}{n}$ for every sample. The number of samples that include a given unit i is $\binom{N-1}{n-1}$. Thus,

$$E(\bar{y}) = \sum \bar{y}_s P(s) = \frac{1}{n} \sum_{i=1}^N y_i \binom{N-1}{n-1} / \binom{N}{n} = \frac{1}{N} \sum_{i=1}^N y_i \quad (2.16)$$

so the sample mean is an unbiased estimator of the population mean under simple random sampling.

Alternatively, the expectation of the sample mean under simple random sampling can be derived using a device that proves useful in many more complicated designs as well. For each unit i in the population, define an indicator variable z_i such that $z_i = 1$ if unit i is included in the sample and $z_i = 0$ otherwise. Then the sample mean can be written in the alternative form

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N y_i z_i \quad (2.17)$$

Each of the z_i is a (Bernoulli) random variable, with expected value $E(z_i) = P(z_i = 1) = n/N$. Hence the expected value of the sample mean is

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N y_i E(z_i) = \frac{1}{n} \sum_{i=1}^N y_i \frac{n}{N} = \frac{1}{N} \sum_{i=1}^N y_i = \mu \quad (2.18)$$

The variance of the sample mean under simple random sampling can be derived similarly by either method. Using the indicator-variable method, the variance is

$$\text{var}(\bar{y}) = \text{var} \left(\frac{1}{n} \sum_{i=1}^N y_i z_i \right) = \frac{1}{n^2} \left[\sum_{i=1}^N y_i^2 \text{var}(z_i) + \sum_{i=1}^N \sum_{j \neq i} y_i y_j \text{cov}(z_i, z_j) \right]$$

Since z_i is a Bernoulli random variable, $\text{var}(z_i) = (n/N)(1 - n/N)$.

The number of samples containing both units i and j , when $i \neq j$, is $\binom{N-2}{n-2}$, so that the probability that both units are included is $\binom{N-2}{n-2} / \binom{N}{n} = n(n-1) / [N(N-1)]$. The product $z_i z_j$ is zero except when both i and j are included in the sample, so

$$E(z_i z_j) = P(z_i = 1, z_j = 1) = \frac{n(n-1)}{N(N-1)}$$

The covariance is

$$\text{cov}(z_i, z_j) = E(z_i z_j) - E(z_i)E(z_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{-n(1-n/N)}{N(N-1)}$$

Thus, the variance of the sample mean is

$$\text{var}(\bar{y}) = \frac{1}{n^2} \left(\frac{n}{N}\right) \left(1 - \frac{n}{N}\right) \left[\sum_{i=1}^N y_i^2 - \frac{1}{N-1} \sum_{i=1}^N \sum_{i \neq j}^N y_i y_j \right]$$

Using the identity

$$\sum_{i=1}^N (y_i - \mu)^2 = \sum_{i=1}^N y_i^2 - \frac{(\sum y_i)^2}{N} = \frac{1}{N} \left[(N-1) \sum_{i=1}^N y_i^2 - \sum_{i=1}^N \sum_{j \neq i}^N y_i y_j \right]$$

the variance expression simplifies to

$$\text{var}(\bar{y}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{\sum (y_i - \mu)^2}{N-1} = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$$

For simple random sampling with replacement, the expected value and variance of the sample mean and the expected value of the sample variance are obtained from the usual statistical properties of the sample mean of independent and identically distributed random variables. On any draw, unit i has probability $p_i = 1/N$ of being selected. The probability that unit i is included (one or more times) in the sample is $\pi_i = 1 - (1 - N^{-1})^n$. The expected number of times unit i is included in the sample is n/N .

2.7. MODEL-BASED APPROACH TO SAMPLING

In the fixed-population or design-based approach to sampling, the values y_1, y_2, \dots, y_N of the variable of interest in the population are considered as fixed but unknown constants. Randomness or probability enters the problem only through the deliberately imposed design by which the sample of units to observe is selected. In the design-based approach, with a design such as simple random sampling the sample mean is a random variable only because it varies from sample to sample. One sample gives a value of the sample mean that is greater than the population mean; another sample gives a value of the sample mean that is lower than the population mean.

In the stochastic-population or model-based approach to sampling, the values of the variable of interest, denoted Y_1, Y_2, \dots, Y_N , are considered to be random variables. The population model is given by the joint probability distribution or density

function $f(y_1, y_2, \dots, y_N; \theta)$, which may depend on one or more unknown parameters θ . The population values y_1, y_2, \dots, y_N realized represent just one outcome of many possible outcomes under the model for the population.

Suppose that the object is to estimate the population mean: for example, mean household expenditure for a given month in a geographical region. Economic theory may suggest a statistical model, such as a normal or lognormal distribution, for the amount a household might spend. The amount that the household spends that month is then one realization among the many possible under the assumed distribution.

As a very simple population model, assume that the population variables Y_1, Y_2, \dots, Y_N are independent, identically distributed (i.i.d.) random variables from a distribution having a mean θ and a variance γ^2 . That is, for any unit i , the variable of interest Y_i is a random variable with expected value $E(Y_i) = \theta$ and variance $\text{var}(Y_i) = \gamma^2$, and for any two units i and j , the variables Y_i and Y_j are independent.

Suppose that we have a sample s of n distinct units from the population and the object is to estimate the parameter θ of the distribution from which the population comes. For the given sample s , the sample mean

$$\bar{Y} = \frac{1}{n} \sum_{i \in s} Y_i$$

is a random variable, whether or not the sample is selected at random, because for each unit i in the sample Y_i is a random variable that can take on different outcomes. With the assumed model, the expected value of the sample mean is $E(\bar{Y}) = \theta$ and its variance is $\text{var}(\bar{Y}) = \gamma^2/n$. Thus, \bar{Y} is a model-unbiased estimator of the parameter θ . An approximate $1 - \alpha$ confidence interval for the parameter θ , based on the central limit theorem for the sample mean of independent, identically distributed random variables, is given by

$$\bar{Y} \pm tS/\sqrt{n}$$

where S is the sample standard deviation and t is the upper $\alpha/2$ point of the t distribution with $n - 1$ degrees of freedom. If additionally the Y_i are assumed to have a normal distribution, then the confidence level is exact, even with a small sample size.

In the study of household expenditure the focus of interest may not be on the parameter θ of the model, however, but on the actual average amount spent by households in the community that month. That is, the object is to estimate (or predict) the value of the random quantity

$$Z = \frac{1}{N} \sum_{i=1}^N Y_i$$

The difference between inference about the random variable Z and the model parameter θ can be appreciated by considering a survey in which every household in a community is included in the sample, so that $n = N$. Then, with the

expenditure Y_i measured for every household, there is no uncertainty about the value of the population mean $Z = (1/N) \sum_{i=1}^N Y_i$. However, even with the whole population observed, there is still uncertainty about the parameter θ of the model that produced the population values, since we have observed only one realization of the N values from that distribution. That is, with the entire population of households observed, there is no uncertainty about the household expenditure realized in that population (assuming no measurement error), but there is uncertainty about the exact distribution or process that produced the expenditure pattern realized. In reality, the more common situation is that the sample size is much smaller than the population size, so that there is uncertainty about both the population values realized and the parameters of their distribution.

To estimate or predict the value of the random variable $Z = (1/N) \sum_{i=1}^N Y_i$ from the sample observations, an intuitively reasonable choice is again the sample mean $\hat{Z} = \bar{Y} = \sum_{i \in s} Y_i / n$. Both Z and \hat{Z} have expected value θ , since the expected value of each of the Y_i is θ . Because $E(\hat{Z}) = E(Z)$, with the expectations evaluated under the assumed model distribution, the predictor \hat{Z} is said to be “model unbiased” for the population quantity Z . More precisely, a predictor \hat{Z} is said to be *model unbiased* for Z if, for any given sample s , the conditional expectations are equal, that is,

$$E(\hat{Z}|s) = E(Z|s)$$

Additionally, for the type of designs we are considering, the expectation of the population quantity Z does not depend on the sample s selected, so that $E(Z|s) = E(Z)$.

Note that the design unbiasedness of the sample mean for the population mean under our assumed model does not depend on how the sample was selected, that is, does not depend on the design. Under the assumed model, the predictor is unbiased with the specific sample selected.

In estimating or predicting the value of a random variable Z with a predictor \hat{Z} , one measure of the uncertainty is the mean square prediction error

$$E(\hat{Z} - Z)^2$$

If the predictor \hat{Z} is model unbiased for Z , then $E(\hat{Z} - Z) = 0$ and the mean square prediction error is the variance of the difference,

$$E(\hat{Z} - Z)^2 = \text{var}(\hat{Z} - Z)$$

In the case of the sample mean $\hat{Z} = \bar{Y}$ as a predictor of the population mean $Z = \sum_{i=1}^N Y_i / N$, with the model in which Y_1, \dots, Y_N are independent, identically distributed from a distribution with mean θ and variance γ^2 , the mean square prediction error is

$$E(\bar{Y} - Z)^2 = \left(\frac{N - n}{N} \right) \frac{\gamma^2}{n}$$

Proof: Because $E(\bar{Y}) = E(Z)$,

$$E(\bar{Y} - Z)^2 = \text{var} \left(\frac{1}{n} \sum_{i \in s} Y_i - \frac{1}{N} \sum_{i=1}^N Y_i \right)$$

Separating the terms for units in the sample s from the units in \bar{s} outside the sample yields

$$\text{var} \left(\bar{Y} - \frac{1}{N} \sum_{i=1}^N Y_i \right) = \text{var} \left[\left(\frac{1}{n} - \frac{1}{N} \right) \sum_{i \in s} Y_i - \frac{1}{N} \sum_{i \in \bar{s}} Y_i \right]$$

Since the values in the sample are independent of the values outside the sample, the variance of the difference between the two independent terms is

$$\begin{aligned} \text{var} \left[\left(\frac{1}{n} - \frac{1}{N} \right) \sum_{i \in s} Y_i - \frac{1}{N} \sum_{i \in \bar{s}} Y_i \right] &= \left(\frac{1}{n} - \frac{1}{N} \right)^2 n \gamma^2 + \frac{1}{N^2} (N - n) \gamma^2 \\ &= \left[\frac{(N - n)^2}{n N^2} + \frac{n(N - n)}{n N^2} \right] \gamma^2 \\ &= \frac{N - n}{n N} \gamma^2 \end{aligned}$$

□

Notice that in the model framework, the finite-population variance

$$V = \frac{\sum_{i=1}^N (Y_i - Z)^2}{N - 1}$$

is itself a random variable. The notation Z in place of μ and V in place of σ^2 is used only to emphasize the model-based viewpoint in which these population quantities are themselves random variables. With the independent, identically distributed model,

$$E(V) = \gamma^2$$

since by standard results in statistics, the expectation of a sample variance of i.i.d. random variables equals the variance of the distribution from which the variables come.

An unbiased estimator or predictor of the mean square prediction error is

$$\hat{E}(\hat{Z} - Z)^2 = \frac{N - n}{N} \frac{S^2}{n}$$

since $E(S^2) = \gamma^2$ since the Y_i are i.i.d. with variance γ^2 .

Further, an approximate $1 - \alpha$ prediction interval for Z is given by

$$\bar{Y} \pm t \sqrt{\hat{E}(\hat{Z} - Z)^2}$$

where t is the upper $\alpha/2$ point of the t distribution with $n - 1$ degrees of freedom. If, additionally, the distribution of the Y_i is assumed to be normal, the confidence level is exact.

Thus, with the assumed i.i.d. model, the estimation and assessment of uncertainty are carried out using exactly the same calculations from the sample data as those used with simple random sampling in the design-based approach. The validity of the inference in the model-based approach does not require that the sample be selected by random sampling, but does depend on the realism of the assumed model.

2.8. COMPUTING NOTES

Simple computations for sampling will be illustrated using the open-source statistical programming language R (Ihaka and Gentleman, 1996). For information about R, the main R Project Web page is <http://www.r-project.org/>. Following the link to Manuals a basic guide is “An Introduction to R.” A good starting point in that manual is the Appendix called “A Sample Session.” Below the official manuals is a link to “contributed” documents, which provides introductions and references ranging in length from one or two pages to several hundred pages. It is generally very quick and easy to install R on your own computer. You can download it from the R-project site. Follow the instructions for Windows, Mac OSX, or Linux, depending on which type of computer system you have.

Entering Data in R

To set the variable x to one specific value:

```
x <- 20
```

To set x to a set of values, say, 3, 7, 5, and 2, use the combine function “c”:

```
x <- c(3, 7, 5, 2)
```

This combines the five values into a vector named “x”.

Another way is to type “scan()” and at the prompt enter the values with spaces between them, rather than commas. Hit enter twice when you are done.

Check that x is what you want by typing either “x” or “print(x)”.

To read data into R from a spreadsheet program such as Excel, the easiest way is to save the data while in the spreadsheet program as a text file with values separated by commas. Then the file can be read into R with the read.table command or the more specialized read.csv command. The details will be provided later whenever we need one of these functions.

Here are some simple R commands to get started:

```
# everything after "#" is a comment line. You don't need to
  type these lines.
```

```
# type the following commands into the R command window.
# produce the x and y coordinates for a randomly
#   distributed population of objects in space (eg, trees
#   or animals):

popnx <- runif(100)
popny <- runif(100)

# plot the spatial distribution of the population
plot(popnx,popny)

# change the size of the circle representing each object:
# plot(popnx,popny,cex=2)

# select a random sample, without replacement, of 10 objects
#   out of the 100 in the population oursample
oursample <- sample(1:100,10)

# draw the sample points, in the same plot
points(popnx[oursample],popny[oursample])

# distinguish the sample points from the others by color
points(popnx[oursample],popny[oursample],
       pch=21,bg="red",cex=2)
```

Sample Estimates

The caribou data of Example 2.1 can be entered and stored as a vector called “y” as follows.

```
y <- c(1, 50, 21, 98, 2, 36, 4, 29, 7, 15, 86, 10, 21, 5, 4)
```

The sample mean is

```
mean(y)
```

Expanding that by the population size gives the estimate of the population total.

```
N <- 286
N * mean(y)
```

The calculations for Example 2.1 along with the output looks like this:

```
> y <- c(1, 50, 21, 98, 2, 36, 4, 29, 7, 15, 86, 10, 21, 5, 4)
# the sample mean:
> mean(y)
[1] 25.93333
# the sample variance:
> var(y)
```

```

[1] 919.0667
# the estimate of the variance of the sample mean:
> (1-15/286) *var(y)/15
[1] 58.05759
# and standard error:
> sqrt(58.06)
[1] 7.619711
# the estimate of the population total:
> 286*25.9333
[1] 7416.924
# the estimated variance of that estimate:
> 286^2 * 58.0576
[1] 4748879
# and standard error:
> sqrt(4748879)
[1] 2179.192
>

```

Simulation

The effectiveness of a sampling strategy can be studied through the use of stochastic simulation. In this method a “population” of N units with corresponding y -values, as similar as possible to the type to be studied, is obtained or constructed by some means and stored in the computer. Then (i) a sample is selected using the design under consideration, such as simple random sampling with sample size n ; and (ii) with the sample data, an estimate of a population characteristic is obtained. These two steps are repeated b times, where the number of runs b is a large number. The b repetitions of the sampling procedure produce b different samples s , each of n units, and b corresponding values of the estimate. The average of these values approximates the expected value of the estimator under the design. The mean square error of the b values approximates the mean square error of the estimator under the design. With an unbiased strategy, the mean square error is the same as the variance.

The sample mean is used as an estimate of the population mean. For a population a data set included in R called “trees” is used. The “tree” data set contains measurements on 31 fir trees, so $N = 31$ in the simulation. A simple random sample of $n = 10$ of these units is selected using the R function “sample” and the sample mean is calculated for the n units in the sample using the R function “mean.” Then a simulation is carried out by repeating the selection and estimation procedure b times using a loop and storing the results in a vector called *ybar*. First $b = 6$ is used to make sure the simulation is working. Then $b = 10,000$ iterations or runs of the simulation are done. Resulting summary statistics are printed and a histogram is plotted showing the distribution of \bar{y} over the many samples.

Lines after the symbol “#” are comments, and are included only for explanation.

```

# Print the trees data set.
trees

```

```
# The variable of interest is tree volume, which for
# simplicity we name "y".
y <- trees$Volume
# The 31 trees will serve as our "population" for
# purposes of simulation.
N <- 31
# sample size:
n <- 10
# Select a simple random sample of n units from 1, 2,..., N.
s <- sample(1:N, n)
# Print out unit numbers for the 10 trees in the sample:
s
# Print the y-values (volumes) of the sample trees:
y[s]
# The sample mean:
mean(y[s])
# Select another sample from the population and repeat the
# estimation procedure:
s <- sample(1:N, n)
s
mean(y[s])
# Compare the estimate to the population mean:
mu <- mean(y)
mu
# Try a simulation of 6 runs and print out the six
# values of the estimate obtained, mainly to check
# that the simulation procedure
# has no errors:
b <- 6
# Let R know that the variable ybar is a vector:
ybar <- numeric(6)
for (k in 1:b){
  s <- sample(1:N,n)
  ybar[k] <- mean(y[s])
}
ybar
# Now do a full-size simulation of 10,000 runs:
b <- 10000
for (k in 1:b){
  s <- sample(1:N,n)
  ybar[k] <- mean(y[s])
}
# Summarize the properties of the sampling strategy
# graphically and numerically:
hist(ybar)
mean(ybar)
var(ybar)
# Compare the variance calculated directly from the
# simulation above to the formula that applies
```

```
# specifically to simple random sampling with the
#v sample mean:
(1-n/N)*var(y)/n
sd(ybar)
sqrt((1-n/N)*var(y)/n)
# The mean square error approximated from the
# simulation should be # close to the variance but
# not exactly equal, since they are calculated
# slightly differently:mean((ybar - mu)^2)

# quit R.
q()
```

The above code shows only the commands typed but not the resulting output. The commands together with the output as shown in the R window are shown below.

```
> y <- trees$Volume
> N <- 31
> n <- 10
> s <- sample(1:N, n)
> s
[1] 4 16 17 7 1 10 25 20 2 22
> y[s]
[1] 16.4 22.2 33.8 15.6 10.3 19.9 42.6 24.9 10.3 31.7
> mean(y[s])
[1] 22.77
> s <- sample(1:N, n)
```

Figure 2.3 shows the histogram obtained in the above simulation with 10,000 runs. The histogram provides a close approximation of the sampling distribution of the estimator \bar{y} with simple random sampling. The actual sampling distribution is over all possible samples, rather than just 10,000. Since the strategy is unbiased the histogram should balance on the actual population mean of the $N = 31$ tree volumes. The variance $\text{var}(\bar{y})$ or standard deviation $\sqrt{\text{var}(\bar{y})}$ summarize the spread of that sampling distribution. If the spread is small, then with high probability the sampling strategy will produce an estimate close to the actual population mean, a desirable property.

```
> s
[1] 1 27 22 11 18 31 8 3 9 20
> mean(y[s])
[1] 30.22
> mu <- mean(y)
> mu
[1] 30.17097
> b <- 6
> ybar <- numeric(6)
```

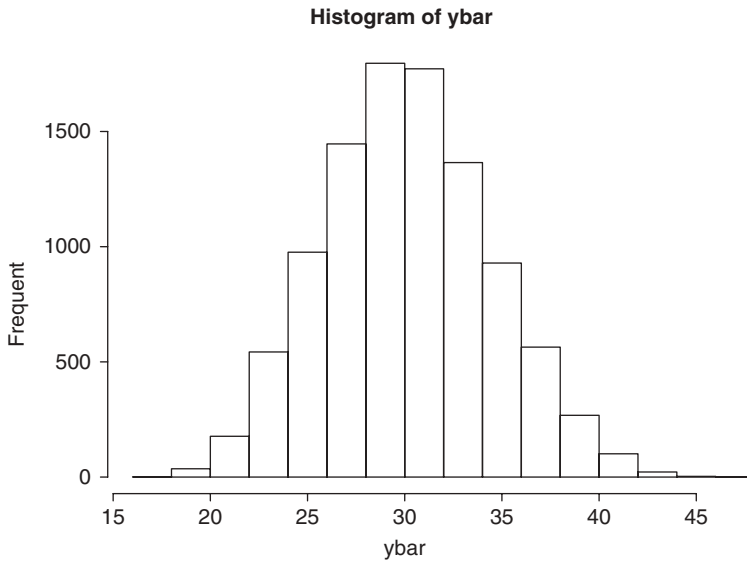


Figure 2.3. Simulation in R of the tree data with $N = 31$, $n = 10$, and $b = 10,000$ iterations.

```
> for (k in 1:b){
+ s <- sample(1:N,n)
+ ybar[k] <- mean(y[s])
+ }
> ybar
[1] 34.08 34.26 31.87 27.45 27.91 31.93
> b <- 10000
> for (k in 1:b){
+ s <- sample(1:N,n)
+ ybar[k] <- mean(y[s])
+ }

>
> hist(ybar)

> mean(ybar)
[1] 30.21421
> var(ybar)
[1] 18.59638
> (1-n/N)*var(y)/n
[1] 18.30406
> sd(ybar)
[1] 4.312352
> sqrt((1-n/N)*var(y)/n)
[1] 4.278324
> mean((ybar - mu)^2)
[1] 18.59639
```

Further Comments on the Use of Simulation

When a researcher conducts a survey, he/she selects a sample of n units from the real population and observes the variables of interest in the sample units. From those data, the researcher calculates estimates of population quantities. For practical reasons including time and cost, it is generally not possible to select more than one sample from the population, nor is it possible to observe the whole population in order to determine how well a certain sampling strategy has done. The best way to see how a given sampling strategy does for a given type of study is to set up a simulation of the sampling situation. Furthermore, in learning about sampling, many people feel that they truly understand it for the first time once they have carried out a simple simulation of a sampling strategy. Since we do not have data on our entire real population, the best we can do is to find or create a population as similar to it as possible and try sampling that population with our strategy again and again to see how it works.

You could do this by having a field study area or other test population that is more accessible, smaller, or more thoroughly studied than the actual population you are interested in, and physically implementing different sampling strategies in it to see how the estimates compare to the known population values there. Generally, though, it is more practical to set up a test population on the computer that is as realistic as you can make it, select a sample from that population with the design you are considering, make estimates using the data from that sample, and compare those estimates with the known characteristics of the test population. Repeat this whole procedure many times, each time selecting a sample of n units with the given design and making estimates using the data from that sample.

Suppose the simulation has 10,000 runs of this type. The distribution of an estimate over those 10,000 trials gives a pretty good picture of the sampling distribution of that estimate with that sampling design for a population of the type used in the simulation.

Example 4: Simulation of simple random sampling using the fir seedling data as a test population. The fir data from the R library “boot” has fir seedling counts for 50 plots, so our test population has $N = 50$ units. We will simulate the sampling strategy simple random sampling design with sample size $n = 5$, using the sample mean as an estimate of the population mean.

In R, load the boot package:

```
library(boot)
```

Take a look at the test population:

```
fir
```

The variable of interest is called “count”. For convenience, call it “y”:

```
y <- fir$count
```

Take a look at `y` for the whole population and get a summary table of its values:

```
y  
table(y)
```

Determine the actual population mean. Note that in the field with your real population you would not have the privilege of seeing this.

```
mean(y)
```

Select a random sample without replacement (simple random sample) of 5 units from the unit labels numbered 1 through 50. A unit label corresponds to a row in the `fir` data file.

```
s <- sample(1:50,5)
```

Look at the unit labels of the selected sample, then at the `y`-values associated with those units. Make sure you recognize the meaning of these by finding them in the `fir` population file you printed out above.

```
s  
y[s]
```

Now compute the sample mean for just the 5 seedling counts in your sample data. Typically, this will not be equal to the population mean, so there is an error associated with the estimate.

```
mean(y[s])
```

Two or more commands like this can be put together on a single line, separated by a semicolon. Repeat the line several times using the up-arrow key and you will see different sample means with different samples.

```
s <- sample(1:50,5);mean(y[s])
```

We are already doing a simulation of the sampling strategy. Each time you enter the above command, a simple random sample is selected and an estimate is made from the sample data.

So far we have not been saving the values of the sample mean for the different samples. We will save them in a vector we'll call "`ybar`." Before starting we tell R to expect `ybar` to be a vector, though without specifying how long it will be.

```
ybar <- numeric()
```


We can put our commands in a “for” loop to get ten simulation runs automatically. This says, as the run number k goes from 1, 2,... to 10, for each run do all the commands within the brackets.

```
for (k in 1:10){s<-sample(1:50,5);ybar[k]<-mean(y[s])}
```

On each run, a simple random sample of 5 units is selected from the test population using simple random sampling. The sample mean for the sample data in that run is stored as the k th component of the vector `ybar`.

Since we only did 10 runs, you can look at the sample mean for each run:

```
ybar
```

Now look at the average of those 10 sample means. Probably it is not the same as the actual population mean, but might be fairly close.

```
mean(ybar)
```

Since we only did 10 runs, we do not get a very good picture of the sampling distribution of the sample mean with our simple random sampling design.

```
hist(ybar)
```

We should get a much better picture of the sampling distribution of our estimator by doing 1000 or more simulation runs:

```
ybar <- numeric(1000)
for (i in 1:1000){s<-sample(1:50,5);ybar[i]<-mean(y[s])}
mean(ybar)
hist(ybar)
```

If you are using a fast computer, you could try it with 10,000 or 100,000 runs, which should result in a nice smooth histogram. (But don't wait forever. If it is taking too long try “control C” or “escape” to stop it.)

With a large number of runs, the average value of the estimator over all the runs should be very close to its theoretically expected value. One can assess the bias or expected error by looking at the difference between the average value of the estimator over all the samples and the true test population characteristic.

Calling the test population mean “ μ ”,

```
mu <- mean(y)
```

the bias in the strategy revealed by the simulation is

```
mean(ybar)-mu
```

With simple random sampling, the sample mean is an unbiased estimator of the population mean, so the above difference should be very close to zero. It would

approach exactly zero if we increased the number of simulation runs. Another common measure of how well a sampling strategy is doing is the mean square error. The simulation value of this quantity is

```
mean((ybar - mu)^2)
```

A good strategy has small (or no) bias and small mean square error.

A simulation study such as this offers usually the most practical approach to assessing how well a certain sampling strategy will work for your study, whether the sample size is adequate, and whether a different sampling design or choice of estimator would work better. In addition, it offers perhaps the best insight into how sampling works and what makes an effective design. \square

EXERCISES

1. In Figure 2.4, the locations of objects (e.g., trees, mines, dwellings) in a study region are given by the centers of “+” symbols. The goal is to estimate the number of objects in the study region.
 - (a) A random sample without replacement of $n = 10$ units has been selected from the $N = 100$ units in the population. Units selected are indicated by

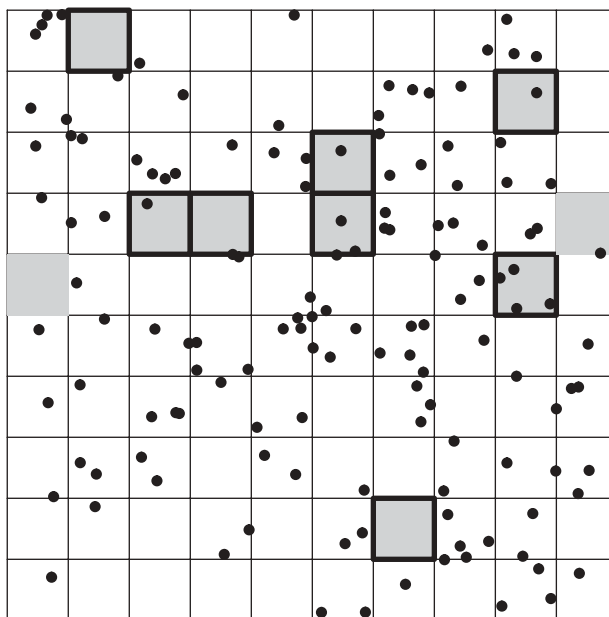


Figure 2.4. Simple random sample of 10 units from a population of 100 units. The variable of interest is the number of point objects within each unit. (See Exercise 1.)

- shading in Figure 2.4. List the sample data. Use the sample to estimate the number of objects in the figure. Estimate the variance of your estimator.
- (b) Repeat part (a), selecting another sample of size 10 by simple random sampling (without replacement) and making new estimates. Indicate the positions of the units of the samples on the sketch.
 - (c) Give the inclusion probability for the unit in the upper left-hand corner. How many possible samples are there? What is the probability of selecting the sample you obtained in part (a)?
2. A simple random sample of 10 households is selected from a population of 100 households. The numbers of people in the sample households are 2, 5, 1, 4, 4, 3, 2, 5, 2, 3.
 - (a) Estimate the total number of people in the population. Estimate the variance of your estimator.
 - (b) Estimate the mean number of people per household and estimate the variance of that estimator.
 3. Consider a small population of $N = 5$ units, labeled 1, 2, 3, 4, 5, with respective y -values 3, 1, 0, 1, 5. Consider a simple random sampling design with a sample size $n = 3$. For your convenience, several parts of the following may be combined into a single table.
 - (a) Give the values of the population parameters μ , τ , and σ^2 . List every possible sample of size $n = 3$. For each sample, what is the probability that it is the one selected?
 - (b) For each sample, compute the sample mean \bar{y} and the sample median m . Demonstrate that the sample mean is unbiased for the population mean and determine whether the sample median is unbiased for the population median.
 4. Show that $E(s^2) = \sigma^2$ in simple random sampling, where the sample variance s^2 is defined with $n - 1$ in the denominator and the population variance σ^2 is defined with $N - 1$ in the denominator. [Hint: Write $y_i - \bar{y}$ as $y_i - \mu - (\bar{y} - \mu)$, verify that

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \mu)^2 - n(\bar{y} - \mu)^2$$
 and either take expectation over all possible samples or define an indicator variable for each unit, indicating whether it is included in the sample.]
 5. The best way to gain understanding of a sampling and estimation method is to carry it out on some real population of interest to you. If you are not already involved in such a project professionally, choose a population and set out to

estimate the mean or total by taking a simple random sample. Examples include estimating the number of trees on a university campus by conceptually dividing the campus into plots, estimating the number of houses in a geographic area by selecting a simple random sample of blocks, or estimating the mean number of people per vehicle during rush hour. In the process of carrying out the survey and making the estimates, think about or discuss with others the following:

- (a) What practical problems arise in establishing a frame, such as a map or list of units, from which to select the sample?
 - (b) How is the sample selection actually carried out?
 - (c) What special problems arise in observing the units selected?
 - (d) Estimate the population mean and total.
 - (e) Estimate the variance of the estimators used in part (d).
 - (f) How would you improve the survey procedure if you were to do it again?
6. Repeat the simulation exercise with the tree data.
- (a) Using $n = 10$ check to see that your results are the same as those in the text.
 - (b) Using $n = 15$ compare your results to those of part (a) first by inspection of the two histograms (note the scale) and by comparing the numerical results summarizing mean and variance of the estimator.