1. From the "Sampling Problems" on pp. 6-7 of the notes, give the sampling unit, population, sampling plan for each stage of sampling for problem 1, all parts. Be specific in your answers.

A researcher has a list of all 4-year colleges and universities in the United States.

(a) The parameter of interest is the proportion of 4-year schools which offer a degree in education. The names of 50 schools are drawn at random from the list and the proportion of these 50 offering such a degree is computed.

Solution:

The sampling unit is a school (4-year college or university) with the population being all such schools in the United States. This is a one-stage simple random sample (SRS).

□

(b) As in (a), but the list is divided into three groups according to enrollment: less than 2000 students, 2000 to 10000 students, greater than 10000 students. Twenty schools are drawn at random from each group.

Solution:

This is a stratified two-stage sampling plan. At the first stage, the sampling unit is school-size-group with the population being each of the three groups and the plan is a census. At the second stage, the sampling unit is a school with the population being schools in a given size-group and the plan is a SRS.

□

(c) The groups in (b) are each subdivided into two groups: those which offer graduate degrees (in any field) and those which do not. Ten schools are drawn at random from each of these subgroups.

Solution:

This adds an intermediate stage to the two in (b). The first stage is the same, the second stage has binary sampling unit of graduate-degree-granting-status given a school-size group and the plan is again a census. In the third stage, the sampling unit is a school, and the population is the collection of schools with a given school-size group and a graduate-degree-granting status. The third stage plan is again a SRS.

□

(d) The parameter of interest is the average age of full-time students in all the schools. Fifty schools are drawn at random and the average age of all students at these 50 schools is computed.

Solution:

This is a two-stage cluster sample. At the first stage, the sampling unit is a school being drawn from the population of all schools in the US, and the plan is a SRS. At the second stage, the sampling unit is a student from the sampling population of all students at a given school, and it is a census.

□

(e) As in (d), except that 100 students are chosen at random from each of the 50 schools and the average age of these students computed.

Solution:

This is a two-stage plan. The first stage is as before and the sampling unit and sampling population are the same as before, but the second stage plan is a SRS.

□

2. Consider a population of size $N = 5$ divided into two strata where the response (y) values for the first stratum are 3, 7, and 8 and for the second stratum are 12 and 15. A stratified random sample consisting of one observation from each stratum will be taken. Let y_1 denote the sample observation from the first stratum and y_2 the sample observation from the second stratum.

In the following table we summarize all possible samples and the corresponding statistics for the following problems. In the bottom row, we calculate the expected value for each estimator.

y_1	y_2	$\overline{y}$	$\overline{y}_s$
3	12	7.5	6.6
7	12	9.5	9.0
8	12	10.0	9.6
3	15	9.0	7.8
7	15	11.0	10.2
8	15	11.5	10.8
		9.75	9

(a) Let $\overline{y} = \frac{1}{2}(y_1 + y_2)$. Derive the sampling distribution of $\overline{y}$ and show that it is a biased estimator of the population mean μ.

Solution:

Note that the mean of the population is $\mu = \frac{1}{5}(3 + 7 + 8 + 12 + 15) = 9$. Yet, the expected value of the estimator $\overline{y}$ is 9.75, hence the estimator is biased in this case.

□

(b) Let $\overline{y}_s = (3/5)y_1 + (2/5)y_2$. Derive the sampling distribution of $\overline{y}_s$ and show that it is an unbiased estimator of μ.

Solution:

On the other hand, the expected value of $\overline{y}_s$ is 9, so it is unbiased for estimating the population mean.

□

(c) Compute the inclusion probability π_1 for each observation in the population. [Note: The inclusion probability π_1 for a unit I is defined in Chapter 6 as the probability that unit i is included in the sample. For an SRS of size n from a population size N,$\pi_i = n/N$ for each unit i.]

Solution:

Since each sample is equally likely, we see from the table above that for each value in the first strata there are two samples in which it is included. I.e. each y_1 has $\pi_i = \frac{2}{6} = \frac{1}{3}$. Similarly, each y_2 has $\pi_i = \frac{3}{6} = \frac{1}{2}$.

□

3. In a square 0.1 acre section of a native hay field, ten 3 ft×2 ft plots were randomly selected. Each was covered by a deer proof exclosure. At the end of the season, all vegetation in each plot was clipped at ground level and air dried. The air-dry weights in grams of the vegetation in the ten plots were: $68, 52, 87, 54, 39, 47, 37, 36, 42, 24$.

(a) Estimate the total production (air-dry weight in grams) for the entire 0.1 acre section if deer had been excluded. Obtain the standard error of the estimate and an approximate 90 % confidence interval for the total.

Solution:

Denote each of the sample observations y_i. The section is 66×66 square feet, hence if we assume that the plots are placed regularly in the section, then there are $N = 22 \times 33 = 726$ possible plots. We estimate the total production as

$$\widehat{\tau} = \frac{N}{n} \sum_{i=1}^{10} y_i = 35283.6 \,\text{grams}.$$

An approximate 90 % confidence interval is given by

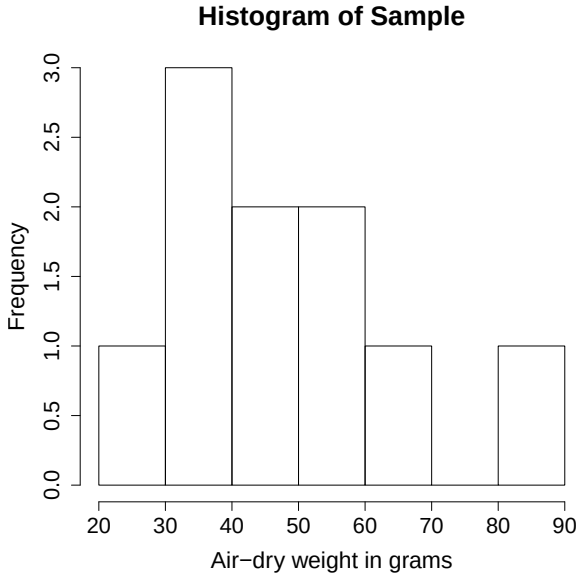
$$\widehat{\tau} \pm t_{.975,9} \text{SE}(\widehat{\tau}) = N\overline{y} \pm t_{.975,9} \sqrt{726(726-10)\frac{s^2}{10}}$$

yields $[25980.2, 44587.0]$.

□

(b) On what assumptions or results is this confidence interval based, and how applicable is the use of the method here?

Solution:



We are assuming that the sample size $n = 10$ and the difference $N - n = 716$ are large enough so that the Finite Central Limit theorem applies and that normal approximation practically approximates the distribution of $\widehat{\tau}$.
The histogram on the left shows that the sample distribution is not wildly skewed, so the convergence is likely sufficient for the scale of our sample.

□

(c) If you were charged with selecting a simple random sample of ten $3' \times 2'$ plots from such a section, how would you do it? Be specific.

Solution:

As mentioned before, there are $N = 726$ possible plots to sample. One could label them on a gridded map in a systematic way (say first by row, then column). Then, using a random number generator that samples from $\{1, 2, \ldots, 726\}$, one could obtain sample of indices, which in turn gives a sample of plots from the section.

□

(d) Estimate how big a sample of plots would be required to estimate the total biomass of the 0.1 acre section to within a margin of error of 3 kg with 95 % confidence.

Solution:

If we think of the previous sample as a pilot study in which we are confident that $\sigma^2 \approx 325.38$, then the minimum sample size required to satisfy $P(|\widehat{\tau} - \tau| > 3000) < .05$ is given by

$$n = \left(\frac{1}{N} + \frac{d^2}{N^2 \sigma^2 z^2} \right)^{-1} = 66.50$$

hence a sample of size of 67 would be sufficient. Based on our confidence of the estimate of σ^2, we may increase this.

□

4. *Thompson 4.1.* A botanical researcher wishes to design a survey to estimate the number of birch trees in a study area. The study area has been divided into 1000 units or plots. From previous experience, the variance in the number of stems per plot is known to be approximately $\sigma^2 \approx 45$. Using simple random sampling, what sample size should be used to estimate the total number of trees in the study area to within 500 trees of the true value with 95 % confidence? To within 1000 trees? To within 2000 trees?

Solution:

d	n	n_0
500	409	692
1000	148	173
2000	42	44

The table summarizes the calculations for both *Thompson 4.1. and 4.2.* The calculation of n was done similarly as in **3.** Each n_0 is the second term in the denominator of calculating n.

□

5. *Thompson 4.2.* Compare the sample sizes for *Thompson 4.1.* when the finite population correction factor is ignored. What do you conclude about the importance of the finite population correction factor for this population?

Solution:

If high precision is necessary in estimating the total, i.e. d is on the order of 500, then the finite population correction factor gives a significantly lower bound than

the naive bound. Although, both cases may not be practical, since they are near a half of the total population and the convergence of the finite central limit theorem might break down. For the less precise schemes, the correction makes less of a difference.

□

6. *Thompson 2.4.* 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. Use the following two strategies. Write $y_i - \overline{y}$ as $y_i - \mu - (\overline{y} - \mu)$, verify that

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \mu)^2 - n(\overline{y} - \mu)^2$$

and compute the expectation in two ways: 1) Take expectation over all possible samples and 2) defining an indicator variable for each unit, indicating whether it is included in the sample and computing the expectation of their sum.

Solution:

To verify the general identity given in the hint, we square then sum both sides of the equality, i.e.

$$(y_i - \overline{y})^2 = (y_i - \mu)^2 - 2(y_i - \mu)(\overline{y} - \mu) + (\overline{y} - \mu)^2$$

so

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \mu)^2 - 2(\overline{y} - \mu)\sum_{i=1}^{n}(y_i - \mu) + n(\overline{y} - \mu)^2 \\ &= \sum_{i=1}^{n}(y_i - \mu)^2 - 2(\overline{y} - \mu)(n\overline{y} - n\mu) + n(\overline{y} - \mu)^2 \\ &= \sum_{i=1}^{n}(y_i - \mu)^2 - n(\overline{y} - \mu)^2.\end{aligned}$$

Now, let us adopt the notation that elements of the population are given by $y_1, y_2, \ldots, y_N$, and for a given sample of indices I of size n, the sample is $y_{I1}, y_{I2}, \ldots, y_{In}$. Further denote the realized sample mean and sample variance of I as

$$\overline{y}_I = \frac{1}{n}\sum_{i=1}^{n} y_{Ii} \quad \text{and} \quad s_I^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{Ii} - \overline{y}_I)^2.$$

In an SRS, each sample is equally likely so $P(I) = 1\Big/\binom{N}{n}$. Then, using the same counting argument in calculating $\mathrm{Var}(\overline{y})$ to rearrange the double summation,

we have

$$\begin{aligned}
E(s^2) &= \sum_{I\in\mathcal{I}} s_I^2 P(I) \\
&= \sum_{I\in\mathcal{I}} \left(\frac{1}{n-1}\sum_{i=1}^{n}(y_{Ii}-\overline{y}_I)^2 \right) \Big/ \binom{N}{n} \\
&= \left(\frac{1}{n-1}\sum_{I\in\mathcal{I}}\sum_{i=1}^{n}(y_{Ii}-\mu)^2 - \frac{n}{n-1}\sum_{I\in\mathcal{I}}(\overline{y}_I-\mu)^2 \right) \Big/ \binom{N}{n} \\
&= \left(\frac{1}{n-1}\binom{N-1}{n-1}\sum_{i=1}^{N}(y_{Ii}-\mu)^2 - \frac{n}{n-1}\binom{N}{n}\mathrm{Var}(\overline{y}) \right) \Big/ \binom{N}{n} \\
&= \frac{1}{n-1}\frac{n}{N}(N-1)\sigma^2 - \frac{\cancel{n}}{n-1}\left(\frac{N-n}{N}\right)\frac{\sigma^2}{\cancel{n}} \\
&= \sigma^2\left(\frac{nN-n-N-n}{(n-1)N}\right) \\
&= \sigma^2.
\end{aligned}$$

Now, let $Z_i \sim \text{Bernouli}(\pi_i)$ indicate whether the population element y_i is in a random sample where the inclusion probability is $\pi_i = \binom{N-1}{n-1}/\binom{N}{n} = \frac{n}{N}$. Using the hint, then the linearity of expectation and that $\mu = E(\overline{y}_I)$ (i.e. it is unbiased), we can write

$$\begin{aligned}
(n-1)E\left(s_I^2\right) &= E\left(\sum_{i=1}^{n}(y_{Ii}-\mu)^2 - n(\overline{y}_I-\mu)^2\right) \\
&= \sum_{i=1}^{N} E\left(Z_i(y_i-\mu)^2\right) - nE\left((\overline{y}_I-\mu)^2\right) \\
&= \sum_{i=1}^{N}\frac{n}{N}(y_i-\mu)^2 - n\mathrm{Var}(\overline{y}_I) \\
&= \frac{n(N-1)\sigma^2}{N} - \cancel{n}\left(\frac{N-n}{N}\right)\frac{\sigma^2}{\cancel{n}} \\
&= \sigma^2\left(\frac{nN-n-N-n}{N}\right) \\
&= (n-1)\sigma^2,
\end{aligned}$$

and canceling $n-1$ on both sides gives the desired equality.

□

7. Suppose you would like to take an SRS of size n from a list of N units, but do not know the population size N in advance. Consider the following procedure:

(a) Set $S_0 = \{1, 2, \ldots, n\}$, so that the initial sample for consideration consists of the first n units on the list.

(b) For $k = 1, 2, \ldots$, generate a random number u_k between 0 and 1. If $u_k > n/(n+k)$, then set S_k equal to S_{k-1}. If $u_k \leq n/(n+k)$, then select one of the units in S_{k-1} at random, and replace it by unit $(n+k)$ to form S_k.

Show that S_{N-n} from this procedure is an SRS of size n. *Hint:* Use induction.

Solution:

Fix the sample size at n, and let $k = N - n$ be the difference in the size between the population to the sample. For clarity, denote $S_k : \Omega \to \{1, 2, \ldots, (n+k)\}^n$ a random variable whose realizations are denoted $S_k(\omega) = s_k$. We must show that S_k takes on each subset of $\{1, 2, \ldots, N = n + k\}$ of size n with equal probability. We proceed by induction on k from $k = 0$ to $k = N - n$.

In the base case, $k = 0$ implies $N = n$ and S_0 takes on only $s_0 = \{1, \ldots, n\}$ with probability 1, which is the only subset of $\{1, \ldots, N = n+0\}$ of size n. Hence it is trivally a random sample of all such subsets.

Now, assume the induction hypothesis, i.e. S_k produces an SRS for a population of size $n + k$ – equivalently,

$$P(S_k = s_k) = 1 \Big/ \binom{n+k}{n}$$

for each subset s_k of size n from $\{1, 2, \ldots, (n+k)\}$. Upon the realization $S_{k+1} = s_{k+1}$, either the sample remains fixed with $S_{k+1} = s_k$, in which case $u_{k+1} \in \left(\frac{n}{n+k+1}, 1\right)$ with probability $\left(1 - \frac{n}{n+k+1}\right) = \left(\frac{k+1}{n+k+1}\right)$, or S_{k+1} takes on s_k with a randomly selected element replaced with $(n+k+1)$, in which case $u_{k+1} \in \left(0, \frac{n}{n+k+1}\right]$ with probability $\left(\frac{k+1}{n+k+1}\right)$. If we assume that the selection of u_{k+1} is independent of the generation of the sample from the previous step, then in the first case

$$\begin{aligned} P(S_{k+1} = s_{k+1}) &= P(S_k = s_k) P\left(U_{k+1} \in \left(\frac{n}{n+k+1}, 1\right)\right) \\ &= \binom{n+k}{n}^{-1} \left(\frac{k+1}{n+k+1}\right) \\ &= \binom{n+k+1}{n}^{-1} \end{aligned}$$

In the second case, denote $s_k = \{a_1, a_2, \ldots, a_n\}$, then $s_{k+1} = s_k \setminus \{a_i\} \cup \{k+1+n\}$ where a_i is the realization of the random selection in the algorithm. For each event where $\{a_1, a_2, \ldots, a_n\} \setminus a_i \subset S_k$, there are $(n+k) - (n-1) = k+1$ possible choices that fix each $a_1, \ldots, a_n$ except for a_i. Since each of those events are disjoint and have equal probability, $P(S_{k+1} = s_{k+1} | u_k, a_i) = P(\{a_1, \ldots, a_n\} \setminus a_i \in S_k) = (k+1)\binom{n+k}{n}^{-1}$. Now, assuming the independence of realization of u_k and a_i and

the previous selection, we have

$$\begin{aligned} P(S_{k+1} = s_{k+1}) &= \left((k+1)\binom{n+k}{n}^{-1} \right) \cdot P\left(U_{k+1} \in \left(0, \frac{n}{n+k+1} \right] \right) \cdot P\Big(A_i = a_i \Big) \\ &= (k+1)\binom{n+k}{n}^{-1} \cdot \frac{n}{n+k+1} \cdot \left(\frac{1}{n} \right) \\ &= \binom{n+k}{n}^{-1} \frac{k+1}{n+k+1} \\ &= \binom{n+k+1}{n}^{-1}. \end{aligned}$$

□