1. Consider a strop adaptive sampling scheme to estimate the number of objects in 10×10 grid below. Label the vertical strips 1 to 10 from left to right. The initial sample will be a systematic sample of two vertical strips with a random starting point from 1 to 5; e.g. , strips 1 and 6 would be one possible sample and strips 4 and 9 would be another, The neighborhood of a plot is the plot itself plus the four plots above, below, and to the right and the left of the plot. If any selected plot has at least one object ,then its neighborhood is included in the sample.

					1				
					14	1			
				1	15	1			
2									
26	1					3	1	1	
3						19	4	25	2
label	1	2	3	4	5	6	7	8	9	10

(a) Calculate the inclusion probabilities for all the non-zero networks. Remember to treat this as a systematic sample of size 1 and not as an SRS of two transects.

Solution

Under the adaptive sampling plan, there are three networks of which we will denote n_1, n_2, and n_3 ordered left to right by their left-most point. Since the sample is of size 1, the inclusion probability α_i is the same as the probability of selection which is given by number systematic samples that include the network over the total number of systematic samples possible. Network n_1 is included in 2 systematic samples, network n_2 is included in 3 samples, and network n_3 is included in 4 samples.

n_i	α_i
n_1	$\frac{2}{5}$
n_2	$\frac{3}{5}$
n_3	$\frac{4}{5}$

□

(b) Derive the sampling distribution of the Horvitz-Thompson estimator of τ, the total number of objects. Show that the estimator is unbiased ($\tau = 120$) and compute its standard deviation.

(c) Derive also the distribution of v across all possible initial samples where v is the

total number of distinct plots which must be surveyed (including edge units). Compute its expected value.

Solution

The Horvitz-Thompson estimator sums over networks in a given sample divided by each respective α_i. Denote y_i the total number of objects in network n_i, then for a given sample s

$$\widehat{\tau} = \sum_{n_i \in s} \frac{y_i}{\alpha_i}.$$

Note also that the number units sample for the ith network, say ν_i, are given respectively $8, 15$, and 13 for $i = 1, 2$, and 3. We summarize this calculation for each of the 5 systematic samples

transects	s	ν	$\widehat{\tau}$
$\{1, 6\}$	$\{n_1, n_2\}$	23	135.00
$\{2, 7\}$	$\{n_1, n_2, n_3\}$	36	203.75
$\{3, 8\}$	$\{n_3\}$	13	68.75
$\{4, 9\}$	$\{n_3\}$	13	68.75
$\{5, 10\}$	$\{n_2, n_3\}$	28	123.75
Mean:		22.6	120.00
Std. Dev.:		8.868	50.0125

□

(d) Calculate the standard deviation of the unbiased estimator of the total for an SRS of $E(v)$ plots (with no adaptive component).

(e) Compare the standard deviation of the H-T estimator in part (b) to the standard deviation of the SRS estimator in part (d). The assumption is that the expected cost of the adaptive plan of part (b) is the same as the cost of the SRS plan of part (d) so that this is a fair comparison of their efficiencies. Is that a valid assumption?

Solution

The standard deviation of an SRS estimator with $n = E(\nu)$ is given by

$$\mathrm{SD}(\tau_{srs}) = \sqrt{N(N-n)\frac{\sigma^2}{n}} \approx 82.93.$$

Assuming that rounding associated with the SRS are insignificant and that the cost is determined only by the number of samples made. Practically, this might not be fair to the adaptive plan, since sampling adjacent cells is likely less expensive than visiting an equal number of randomly selected ones.

□

2. It was desired to estimate the proportion of a large tract of land that could be classified as wetland. The area was first stratified based on aerial photography into three terrain types and then the proportion of wetland estimated within each. In the first stratum, comprising 50 % of the area, a random sample of 100 points was chosen and the points classified: 30 were

classified as wetland. In the second stratum, comprising 30 % of the area, a systematic sample of transects with random starting point was used; the estimated proportion of wetland in this stratum was .42 with SE = .08. In the third stratum, comprising the remaining 20 % of the area, two-stage sample with transects as the primary units and points long the transects as secondary units was used; the estimated proportion of wetland was .10 with SE = .04. Estimate the proportion of the entire area that is wetland and compute an SE for this estimate.

Solution

Denote $\widehat{p}_h$ as the estimated proportion of wetland and w_h the proportion of the study area in stratum $h = 1, 2,$ or 3. We can estimate the total proportion with

$$\widehat{p} = \sum_{h=1}^{3} w_h \widehat{p}_h.$$

Moreover, with separate estimates of the variance of the estimator within each stratum, say $\widehat{\mathrm{Var}}(\widehat{p}_h)$, an estimate for the total variance is given by

$$\widehat{\mathrm{Var}}(\widehat{p}) = \sum_{h=1}^{3} \widehat{\mathrm{Var}}(\widehat{p}_h).$$

We have each of these components given except for the first stratum which is a standard proportion estimate (without finite population). So $p_1 = .3$ and

$$\widehat{\mathrm{Var}}(\widehat{p}_1) = w_1^2 \frac{p_1(1-p_1)}{n_1}.$$

h	w_h	$\widehat{p}_h$	$\mathrm{SD}(\widehat{p}_h)$
1	.5	0.30	0.0229
2	.3	0.42	0.08
3	.2	0.10	0.04
Estimate		0.296	0.0923

□

3. National Marine Fisheries Service places observers on commercial fishing vessels. One of their jobs is to sample the hauls of fish for numbers and species of fish. One method of sampling is basket sampling where baskets of fish are taken from the haul as it is unloaded onto the ship. Suppose one particular haul of fish weights 10,000 kg. Six baskets of fish are selected randomly from the haul. The number and total weight of the fish in each basket are recorded.

(a) How would you estimate the total number of fish in the whole haul and a standard error from the given information? Give the formulas, being careful to specify precisely what each variable represents.

Solution

We can consider each basket a sampling unit and let y_i denote the number of fish in a given basket from a sample of size n. Here we do not know the total number

of baskets, but if we denote x_i the weight of each basket, and $\tau_x = 10,000\,\text{kg}$, we can use a ratio estimator to estimate τ_y, the total number of fish. That is

$$\widehat{\tau}_y = \tau_x r = \tau_x \frac{\overline{y}}{\overline{x}}.$$

where $\overline{y}$ and $\overline{x}$ are the sample means of y_i and x_i respectively.

Depending on the relative fish capacity of the baskets and the number of baskets sampled, we have two possible candidates for estimating the variance (hence standard error) of this estimator. If the amount of fish in a basket is small compared to to total number of fish in the haul, then we would expect that the number baskets sampled is small to the total possible number, and we can ignore the effect of having a finite number of baskets and estimate using the linearized variance of the ratio estimator

$$\widehat{\text{Var}}(\widehat{\tau}_y) = \tau_x^2 \frac{1}{\mu_x^2} \frac{s_r^2}{n} \quad \text{where } s_r^2 = \sum_{i=1}^{n} (y_i - r x_i)^2.$$

If the number of baskets sampled is large, then we can reduce this estimate by multiplying by a finite population correction factor $\widehat{N}(\widehat{N} - n)$, where we estimate the total number of possible baskets by $\widehat{N} = \tau_x / \overline{x}$.

□

(b) The number of salmon in each basket is also recorded. How would you estimate the proportion and SE of all fish in the haul that are salmon? Again, give the formulas, being careful to specify precisely what each variable represents.

Solution

Continuing with the notation from above, we now let z_i be the number salmon in a basket. There are two candidates for an estimate of the proportion. First, for each basket sampled, let $p_i = z_i / y_i$, and use the SRS proportion estimator

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} p_i, \text{ and } \widehat{\text{Var}}(\widehat{p}) = \frac{\widehat{p}(1 - \widehat{p})}{n - 1}$$

where we can add a finite population correction factor $1 - n/\widehat{N}$ reasoning as in (a).

Alternatively, we could use a ratio estimator (with similar fpc reasoning)

$$\widehat{p}_r = \frac{\overline{z}}{\overline{y}}, \text{ with } \widehat{\text{Var}}(\widehat{p}_r) = \frac{1}{\overline{y}^2} \frac{s_r^2}{n} \quad \text{where } s_r^2 = \sum_{i=1}^{n} (z_i - p_r y_i)^2.$$

I suspect that ratio estimation would not give much of a reduction in the standard error from the first method if similarly sized baskets were used since not much auxiliary information is added by considering y_i for each basket. On the other hand, if the baskets were significantly different (i.e. bought at different times from different manufacturers), ratio estimation might provide a better estimate.

□

4. The linearization approximation for the variance of the product of two random variables is $\text{Var}(XY) \approx \mu_Y^2\sigma_X^2 + \mu_X^2\sigma_y^2 + 2\mu_X\mu_Y\rho\sigma_X\sigma_Y$. Suppose X and Y are independent. Derive an exact expression for the variance of XY in terms of the means and variances of X and Y. Note that it is not the same as the linearization approximation with $\rho = 0$.

Solution

Independence gives $\text{E}\,XY = \text{E}\,X\,\text{E}\,Y = \mu_X\mu_Y$, and $\text{E}(XY)^2 = \text{E}(X^2)\,\text{E}(Y^2)$ since $f(X)$ and $f(Y)$ are also independent. Hence

$$\begin{aligned}
\text{Var}(XY) &= \text{E}\left(XY - \text{E}\,XY\right)^2 \\
&= \text{E}\left((XY)^2 - 2XY\,\text{E}\,XY + (\text{E}\,XY)^2\right) \\
&= \text{E}\left((XY)^2 - 2XY\mu_X\mu_Y + (\mu_X\mu_Y)^2\right) \\
&= \text{E}(X^2)\,\text{E}(Y^2) - (\mu_X\mu_Y)^2 \\
&= (\sigma_X^2 + \mu_X^2)(\sigma_Y^2 + \mu_Y^2) - (\mu_X\mu_Y)^2 \\
&= \mu_X^2\sigma_Y^2 + \sigma_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2
\end{aligned}$$

□

5. Suppose I'm interested in the total number of ducklings produced in a large breeding region in North Dakota. I have an estimate, say P, of the total number of ponds in the entire region based on aerial photos. I also have a standard error, $\text{SE}(P)$ for this estimate. In addition, based on ground surveys independent of the aerial phot survey, I have an estimate of the average number of breeding pairs, say B, per pond, along with $\text{SE}(B)$. Finally, from a different study reported in the literature, I have an estimate of the average number of ducklings per pair, say D, along with $\text{SE}(D)$. Estimate the total number of ducklings in the area, along with a standard error for the following data: $P = 23890$, $\text{SE}(P) = 1122$; $B = 2.42, \text{SE}(B) = 0.43$; $D = 5.65, \text{SE}(D) = 0.51$ (Hint: use the linearization approximation given in the previous problem twice. You could also use the exact expression derived in the previous problem since the random variables here are independent.)

Solution

The true number of duckling pairs, say τ, is given by the product of each of the parameters being estimated above. If we assume that each estimator is truely independent, then their product is an unbiased estimator of the total number of ducklings. To obtain an estimate of the standard error of this estimator, we can recursively apply the previous problem to obtain

$$\begin{aligned}
\text{Var}(PDB) &= (\text{E}\,PD)^2\sigma_B^2 + \text{Var}(PD)\sigma_B^2 + \mu_B^2\text{Var}(PD) \\
&= \mu_P^2\mu_D^2\sigma_B^2 + (\mu_P^2\sigma_D^2 + \sigma_P^2\sigma_D^2 + \mu_D^2\sigma_P^2)(\sigma_B^2 + \mu_B^2).
\end{aligned}$$

Hence, replacing each parameter with its respective estimate, we obtain

$$\widehat{\tau} = 3.266 \times 10^5 \text{ ducks} \quad \text{and} \quad \text{SE}(\widehat{\tau}) = 0.671 \times 10^5 \text{ ducks}.$$

□

6. From Chapter 15, derive the expression for the true variance of $\widehat{\tau}$ for simple random sampling with estimated detectability (equation (16.9) of Thompson).

Solution

Given a population of τ objects, we take a SRS of size n from N plots, say y_i, containing those objects. We model imperfect detection of each observed plot y_i, by Bernouli trials $y_i \sim \text{Bernoulli}(Y_i, p)$ where Y_i is the true number of objects in plot i and p is the common probability of detection. In the situation where we must estimate $\widehat{p}$, the estimator for the population total is

$$\widehat{\tau} = \frac{N\overline{y}}{\widehat{p}}.$$

We will derive an estimate for the variance of this estimator. We can linearize the ratio $\overline{y}/\widehat{p}$, so that

$$\frac{1}{N^2}\text{Var}(\widehat{\tau}|S) \approx \frac{1}{(\text{E}\,\widehat{p})^2}\text{Var}(\overline{y}) + \frac{(\text{E}\,\overline{y})^2}{(\text{E}\,\widehat{p})^4}\text{Var}(\widehat{p}) = \frac{1}{p^2}\text{Var}(\overline{y}) + \frac{(\text{E}\,\overline{y})^2}{p^4}\text{Var}(\widehat{p}), \qquad (1)$$

where we have assumed $\widehat{p}$ and $\overline{y}$ to be uncorrelated. Hence, we need only find expressions for $\text{E}\,\overline{y}$ and $\text{Var}(\overline{y})$. Conditioning on a given sample of indices, say S, we have

$$\begin{aligned}
\text{E}\,\overline{y} &= \text{E}\left[\text{E}(\overline{y}|S)\right] \\
&= \text{E}\left[\frac{1}{n}\sum_{i\in S}\text{E}(y_i|S)\right] \\
&= \text{E}\left[\frac{1}{n}\sum_{i\in S}Y_i p\right] \\
&= \mu p,
\end{aligned}$$

$$\begin{aligned}
\text{and } \text{Var}(\overline{y}) &= \text{E}[\text{Var}(\overline{y}|S)] + \text{Var}\left[\frac{1}{n}\sum_{i\in S}Y_i p\right] \\
&= \text{E}\left[\text{Var}\left(\frac{1}{n}\sum_{i\in S}y_i\middle|S\right)\right] + \left(\frac{N-n}{N}\right)\frac{p^2\sigma^2}{n} \\
&= \text{E}\left[\frac{1}{n^2}\sum_{i\in S}Y_i p(1-p)\right] + \left(\frac{N-n}{N}\right)\frac{p^2\sigma^2}{n} \\
&= \frac{p(1-p)}{n}\mu + \left(\frac{N-n}{N}\right)\frac{p^2\sigma^2}{n}.
\end{aligned}$$

Substituting back into (1), gives the desired expression

$$\text{Var}(\widehat{\tau}) = N^2\left[\frac{\mu(1-p)}{np} + \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} + \frac{\mu^2}{p^2}\text{Var}(\widehat{p})\right].$$

□

```
> ##############################################################################
> # Problem 1
> ##############################################################################
> y = c(2,26,3, 1,
+       1, 1,14,15, 1,1,
+       3,19, 1,4, 1,25, 2
+      )
> net = c( rep(1,4),
+          rep(2,6),
+          rep(3,7))
> n = tapply(y,net,sum)
> a = c(2/5,3/5,4/5)
> idx = list()
```

```
> idx[[1]] = c(1,2)
> idx[[2]] = c(1,2,3)
> idx[[3]] = c(3)
> idx[[4]] = c(3)
> idx[[5]] = c(2,3)
> nu.i = c(8,15,13)
> tau.pi = sapply(idx,function(i)sum((n/a)[i]))
> nu     = sapply(idx,function(i)sum(nu.i[i]))
> (mu.tau.pi = mean(tau.pi))
[1] 120

> (sigma.tau.pi = sqrt(sum( (tau.pi - mu.tau.pi)^2 )/5) )
[1] 50.0125

> (mu.nu = mean(nu))
[1] 22.6

> (sigma.nu = sqrt(sum( (nu - mu.nu)^2 )/5) )
[1] 8.86792

> N = 100

> sigma2 = var(c(y,rep(0,N-length(y))))
> (sd.tau.srs = sqrt(N*(N- mu.nu)*sigma2/mu.nu))
[1] 82.92908
> ##############################################################################
> # Problem 2
> ##############################################################################
> w = c(); p = c(); var.p = c();
> w[1] = .5
> p[1] = .3
> var.p[1] = w[1]^2*(p[1]*(1-p[1]))/100
> w[2] = .3
> p[2] = .42
> var.p[2] = .08^2
> w[3] = .2
> p[3] = .1
> var.p[3] = .04^2
> (phat = sum(p*w))
[1] 0.296

> (sdhat = sqrt(sum(var.p)))
[1] 0.09233093
> ##############################################################################
> # Problem 5
> ##############################################################################
> p = 23890
> b = 2.42
> d = 5.65
> se.p = 1122
> se.b = .43
> se.d = .51
> (tauhat = p*b*d)
[1] 326648

> (se.tauhat = sqrt(p^2*d^2*se.b^2 + (p^2*se.d^2 + se.p^2*se.d^2 + d^2*se.p^2)*(se.b^2 + b
    ^2)))
[1] 67158.79
```