Survey Sampling

Statistics 4234/5234 — Fall 2017

Second in-class exam

Answers

1. Consider a population of six students, whose test scores y_i are

Student	1	2	3	4	5	6
Score	66	59	70	83	82	71

The mean and variance of this population are 71.83 and 86.17, respectively.

(a) There are 15 possible SRS's of size 4.

1234	1235	1236	1245	1246
1256	1345	1346	1356	1456
2345	2346	2356	2456	3456

(b) Let stratum 1 consist of students 1–3 and stratum 2 consist of students 4–6. The possible stratifed random samples of size 4, in which 2 students are selected from each stratum, are

1245	1246	1256
1345	1346	1356
2345	2346	2356

(c) The sampling distribution of $\bar{y}_{\text{str}}$ for the stratified sampling scheme described in part (b):

k	69.50	69.75	70.50	70.75	72.25	72.50	73.50	75.25
$\Pr(\bar{y}_{\text{str}} = k)$	1/9	1/9	1/9	1/9	1/9	2/9	1/9	1/9

(d) i. $E(\bar{y}_{\text{str}}) = 71.83 = \bar{y}_U = E(\bar{y})$

ii. $V(\bar{y}_{\text{str}}) = 3.14 < 7.18 = V(\bar{y})$.

2. Norwegian researchers used stratification techniques to estimate ringed seal populations in Svalbard fjords. The study area was divided into three zones, and each zone was divided into a number of plots (N_h in the table below). A random sample of plots in each zone was examined, and the number of breathing holes in each sampled plot was recorded.

Zone	N_h	n_h	$\bar{y}_h$	s_h
1	68	17	1.76	1.82
2	84	12	4.42	3.40
3	48	11	10.55	6.79

(a) Estimate the total number of breathing holes in the study region.

$$\hat{t}_{\text{str}} = \sum N_h \bar{y}_h = 68(1.76) + 84(4.42) + 48(10.55) = 997$$

(b) Give the standard error of your estimate in part (a).

$$\hat{V}(\hat{t}_{\text{str}}) = \sum N_h \frac{s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$$

so

$$\hat{V}(\hat{t}_{\text{str}}) = 68^2 \frac{1.82^2}{17}\left(1 - \frac{17}{68}\right) + 84^2 \frac{3.40^2}{12}\left(1 - \frac{12}{84}\right) + 48^2 \frac{6.79^2}{11}\left(1 - \frac{11}{48}\right)$$

$$= 675.73 + 5826.24 + 7443.72 = 13{,}945.69$$

and thus

$$\text{SE}(\hat{t}_{\text{str}}) = \sqrt{\hat{V}\left(\hat{t}_{\text{str}}\right)} = \sqrt{13{,}945.69} = 118.09$$

(c) You have been asked to design a follow-up survey to estimate the total number of breathing holes at a later date; a total of 40 plots are again to be sampled. How many plots would you sample from each of Zones 1, 2 and 3?

Zone	N_h	s_h	$N_h s_h$	$n_{h,\text{opt}}$
1	68	1.82	123.76	6.73
2	84	3.40	285.60	15.54
3	48	6.79	325.92	17.73
Sum			735.28	40

So take $n_1 = 7$ *and* $n_2 = 15$ *and* $n_3 = 18$.

3. In a simple random sample of $n = 31$ black cherry trees from a forest of $N = 2967$ trees, the mean timber volume per tree was 30.17 cubic feet, with a standard deviation of 16.44 cubic feet.

(a) Calculate the standard error of $\bar{y} = 30.17$ as an estimate of $\bar{y}_U$, the mean timber volume per tree for *all* trees in the forest.

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{16.44^2}{31}\left(1 - \frac{31}{2967}\right) = 8.63$$

and thus

$$\text{SE}(\bar{y}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{8.63} = 2.94$$

(b) Give a 95% confidence interval for the *total* timber volume of all trees in the forest.

CI for $\bar{y}_U$ *is*

$$\bar{y} \pm 1.96 \cdot \text{SE}(\bar{y}) \;\Rightarrow\; 30.17 \pm 5.76 \;\Rightarrow\; [24.41,\ 35.93]$$

and a CI for t_y *is* $N \times (\bar{y} \pm 1.96 \cdot \text{SE}(\bar{y}))$ *with* $N = 2967$, *so*

$$89{,}514 \pm 17{,}081 \;\Rightarrow\; [72{,}433,\; 106{,}595]$$

and thus we are 95% confident that the total volume of black cherry tree timber in this forest is between 72,433 and 106,595 cubic feet.

Now suppose it is known that the sum of the diameters for all the trees in the forest is $t_x = 41{,}835$ inches. For the 31 sampled trees, the average girth (diameter) was 13.25 inches, with a standard deviation of 3.14 inches. The sample correlation between girth and volume was 0.967.

(c) Use ratio estimation to estimate the total volume for all trees in the forest.

$$\hat{t}_{yr} = \hat{B} t_x = \frac{\bar{y}}{\bar{x}} t_x = \frac{30.17}{13.25} \times 41{,}835 = 95{,}258$$

(d) Use regression estimation to estimate the total volume for all trees in the forest.

$$\hat{B}_1 = r\frac{s_y}{s_x} = 0.967\left(\frac{16.44}{3.14}\right) = 5.063$$

so

$$\hat{\bar{y}}_{\text{reg}} = \bar{y} + \hat{B}_1\left(\bar{x}_U - \bar{x}\right) = 30.17 + 5.063\left(\frac{41{,}835}{2967} - 13.25\right) = 34.47$$

and

$$\hat{t}_{y,\text{reg}} = N\hat{\bar{y}}_{\text{reg}} = 2967(34.47) = 102{,}284$$

(e) Suppose the relative efficiency of ratio estimation versus the ordinary sample mean is about 2.5, and the relative efficiency of regression estimation versus ratio estimation is about 6; that is

$$\frac{\hat{V}(\bar{y})}{\hat{V}(\hat{\bar{y}}_r)} = 2.5 \qquad \text{and} \qquad \frac{\hat{V}(\hat{\bar{y}}_r)}{\hat{V}(\hat{\bar{y}}_{\text{reg}})} = 6.0$$

Give the shortest possible (valid) 95% confidence interval for total timber volume you can find from these data.

The most efficient estimator is $\hat{\bar{y}}_{reg}$ with estimated variance $\hat{V}\left(\hat{\bar{y}}_{reg}\right) = \frac{\hat{V}(\bar{y})}{15}$ and thus the narrowest possible interval for t_y is

$$102{,}284 \pm 17{,}081/\sqrt{15} \;\Rightarrow\; 102{,}285 \pm 4410 \;\Rightarrow\; [97{,}874,\; 106{,}695]$$

and we are 95% confident that the total volume of black cherry tree timber in this forest is between 97,874 and 106,695 cubic feet.

4. Suppose we have the resources to take a total sample size of n, and sufficiently precise information about the population that we can intelligently define n strata. We can then take a stratified random sample of just one unit in each stratum. Does this seem like a good idea? What practical or statistical problem might result from such a sampling scheme?

I would not recommend this, since with only one unit per stratum we will have no assessment of the within-strata variability, and thus no standard error for our estimate.