

Topic#4 Selecting the Sample Size for Estimating Population Means and Totals

The amount of information in a sample depends on the sample size n because $V(y_{st})$ decreases as n increases. We now examine a method of choosing the sample size to obtain a fixed amount of information for estimating a population parameter. Suppose we specify that the estimate of y_{st} should lie within B units of the population mean, with probability approximately equal to .95.

$$\text{So... } 2\sqrt{V(y_{st})} = B \quad \text{or} \quad V(y_{st}) = \frac{B^2}{4}$$

This equation contains the actual population variance of y_{st} rather than the estimated variance. The number of observations, n_i allocated to the i th stratum is some fraction of the total sample size n . We denote this fraction by a_i .

$$\text{So... } n_i = na_i$$

Similarly, the estimation of the population total t with a bound of B units on the error of estimation leads to the equation :

$$2\sqrt{V(Ny_{st})} = B \quad \text{or} \quad V(y_{st}) = \frac{B^2}{4N^2}$$

Approximate sample size required to estimate μ or t with a bound of B on the error of estimation:

$$n = \frac{\sum_{i=1}^L N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2}$$

$$D = \frac{B^2}{4} \text{ when estimating } \mu$$

$$D = \frac{B^2}{4N^2} \text{ when estimating } t$$

Example: Lets go back to the example from last week about the estimated television viewing in an area that includes two towns (Town A & B) and a rural area. We would like to estimate the population mean by using y_{st} . The allocation per stratum is equal (given by $a_1 = 1/3$, $a_2 = 1/3$, and $a_3 = 1/3$).

Recall that: $N_1 = 155$, $N_2 = 62$ and $N_3 = 93$

A prior survey suggests that $\sigma_1^2 = 25$, $\sigma_2^2 = 225$, and $\sigma_3^2 = 100$

We would like the bound for error to be 2 hours so $D = \frac{2^2}{4} = 1$

$$\begin{aligned} \sum_{i=1}^3 \frac{N_i^2 \sigma_i^2}{a_i} &= \frac{(155)^2(25)}{(1/3)} + \frac{(62)^2(225)}{(1/3)} + \frac{(93)^2(100)}{(1/3)} \\ &= (24,025)(75) + (3844)(675) + (8649)(300) = 6,991,275 \end{aligned}$$

$$\sum_{i=1}^3 N_i \sigma_i^2 = (155)(25) + (62)(225) + (93)(100) = 27,125$$

$$N^2 D = (310)^2(1) = 96,100$$

$$n = \frac{\sum_{i=1}^L N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2} = \frac{6,991,275}{96,100 + 27,125} = 56.7 \quad \text{or } n = 57$$

$$n_1 = 15\left(\frac{1}{3}\right) = 19 = n_2 = n_3$$

Example: Suppose we would like to estimate the population total in the problem above with a bound of 400 hours on the error of estimation. Choose the appropriate sample size if an equal number of observations is to be taken from each stratum.

$$D = \frac{B^2}{4N^2} = \frac{(400)^2}{4N^2} = \frac{40,000}{N^2} \qquad \sum_{i=1}^3 \frac{N_i^2 \sigma_i^2}{a_i} = 6,991,275 \qquad \sum_{i=1}^3 N_i \sigma_i^2 = 27,125$$

$$n = \frac{\sum_{i=1}^L N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2} = \frac{6,991,275}{N^2 \left(\frac{40,000}{N^2}\right) + 27,125} = 104.2 \text{ or } 105$$

Then $n_1 = n_2 = n_3 = 35$.

Topic#5 Allocation of the Sample

The objective of the sample survey design is to provide estimators with small variances at the lowest possible cost. After a sample size n is chosen, there are many ways to divide n into the individual stratum sample sizes n_1, n_2, \dots, n_L . Each division may result in a different variance for the sample mean. Hence, our objective is to use an allocation that gives a specified amount of information at a minimal cost.

The best allocation scheme is affected by three factors:

1. The total number of elements in each stratum.
2. The variability of observations within each stratum.
3. The cost of obtaining an observation from each stratum.

* The number of elements in each stratum affects the quantity of information in the sample. A sample size 20 from a population of 200 elements should contain more information than a sample of 20 from 20,000 elements. Thus, large-sample sizes should be assigned to strata containing large numbers of elements.

* Variability must be considered because a larger sample is needed to obtain a good estimate of a population parameter when the observations are less homogeneous.

* If the cost of obtaining an observation varies from stratum to stratum, we take small samples from strata with high costs. We do so because our objective is to keep the cost of sampling at a minimum.

The proportion of the observations that should be surveyed for stratum i in order to minimize cost and variance is:

$$\frac{N_i \sigma_i / \sqrt{c_i}}{\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k}} \quad (\text{where } c \text{ is the cost of obtaining an single observation})$$

The total number of observations taken from all the strata that minimizes cost and variance is:

$$\frac{(\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k}) (\sum_{i=1}^L N_i \sigma_i \sqrt{c_i})}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2}$$

Example: Lets go back to the example from last week about the estimated television viewing in an area that includes two towns (Town A & B) and a rural area. We would like to find the overall sample size and the sample size for each strata that would minimize cost and variance. The cost of obtaining an observation from town A or B is estimated to be \$9 (or $c_1 = c_2 = 9$). The cost of obtaining an observation from the rural area is higher due to the cost of traveling from one rural house to another. The cost of one observation in the rural area is estimated to be \$16 (or $c_3 = 16$). Our bound for error will remain 2 hours. Recall that: $N_1 = 155$, $N_2 = 62$ and $N_3 = 93$

A prior survey suggests that $\sigma_1 = 5$, $\sigma_2 = 15$, and $\sigma_3 = 10$

We would like the bound for error to be 2 hours so $D = \frac{2^2}{4} = 1$

We have
$$\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k} = N_1 \sigma_1 / \sqrt{c_1} = N_2 \sigma_2 / \sqrt{c_2} + N_3 \sigma_3 / \sqrt{c_3}$$

$$= \frac{155(5)}{\sqrt{9}} + \frac{62(15)}{\sqrt{9}} + \frac{93(10)}{\sqrt{16}} = 800.83$$

and
$$\sum_{i=1}^L N_i \sigma_i \sqrt{c_i} = N_1 \sigma_1 \sqrt{c_1} + N_2 \sigma_2 \sqrt{c_2} + N_3 \sigma_3 \sqrt{c_3}$$

$$= 155(5)\sqrt{9} + 62(15)\sqrt{9} + 93(10)\sqrt{16} = 8835$$

thus
$$\frac{(\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k})(\sum_{i=1}^L N_i \sigma_i \sqrt{c_i})}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2} = \frac{(800.83)(8835)}{(310)^2(1) + 27,125} = 57.42 = 58$$

then
$$p_1 = \frac{N_1 \sigma_1 / \sqrt{c_1}}{\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k}} = \frac{155(5)/3}{800.83} = 0.32 \quad n_1 = 58(0.32) = 18$$

$$p_2 = \frac{N_2 \sigma_2 / \sqrt{c_2}}{\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k}} = \frac{62(15)/3}{800.83} = 0.39 \quad n_2 = 58(0.39) = 23$$

$$p_3 = \frac{N_3 \sigma_3 / \sqrt{c_3}}{\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k}} = \frac{93(10)/4}{800.83} = 0.29 \quad n_3 = 58(0.29) = 17$$

Hence, the experimenter should select a total of 58 households with 18 from town A, 23 from town B and 17 from the rural area. He or she can then estimate the average number of hours spent watching television at a minimum cost with a bound of 2 hours on the error of estimation.

In some stratified sampling problems, the **cost of obtaining an observation is the same for all the strata**. If we set $c_1 = c_2 = \dots c_L = 1$, then the cost terms cancel out and we get the following two formula.

The proportion of the observations that should be surveyed for stratum i (with equal survey costs) in order to minimize variance is:

$$\frac{N_i \sigma_i}{\sum_{k=1}^L N_k \sigma_k}$$

The total number of observations taken from all the strata equal survey costs that minimizes variance is:

$$\frac{(\sum_{k=1}^L N_k \sigma_k)^2}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2}$$

Example: Lets look at the town A, town B, rural area problem again. Suppose the advertising firm wants to do telephone interviews rather than personal interviews because all the households in the area have telephones, and this method reduces costs. The cost of obtaining an observation is then the same in all three strata. We will keep the 2 hour bound on the error. Find the appropriate sample size n and strata sizes n_1 , n_2 , and n_3 . Recall that: $N_1 = 155$, $N_2 = 62$ and $N_3 = 93$ and $\sigma_1 = 5$, $\sigma_2 = 15$, and $\sigma_3 = 10$

We would like the bound for error to be 2 hours so $D = \frac{2^2}{4} = 1$

$$\sum_{i=1}^L N_k \sigma_k = N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3 = (155)(5) + (62)(15) + (93)(10) = 2635$$

$$\frac{(\sum_{k=1}^L N_k \sigma_k)^2}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2} = \frac{(2635)^2}{96,100 + 27,125} = 56.34 = 57$$

$$p_1 = \frac{N_1 \sigma_1}{\sum_{k=1}^L N_k \sigma_k} = \frac{(155)(5)}{2635} = 0.30, \quad n_1 = 57(0.30) = 17$$

$$p_2 = \frac{N_2 \sigma_2}{\sum_{k=1}^L N_k \sigma_k} = \frac{(62)(15)}{2635} = 0.35, \quad n_2 = 57(0.35) = 20$$

$$p_3 = \frac{N_3 \sigma_3}{\sum_{k=1}^L N_k \sigma_k} = \frac{(93)(10)}{2635} = 0.35, \quad n_3 = 57(0.35) = 20$$

When you compare this example(equal costs) to the previous example, the sample size is almost the same, but the allocation has changed. More observations are taken from the rural area because these observations no longer have a higher cost.