Sample errors and sampling bias

---

8 marks

Consider a population $\mathcal{P}$ of N units $u \in \mathcal{P}$ and suppose that on each unit we have the value of the variate $y_u = y(u)\ \forall u \in \mathcal{P}$. Without loss of generality, we will take $u = 1, \ldots, N$ and denote the ordered variate values in the population $\mathcal{P}$ by

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(N-1)} \leq y_{(N)}.$$

Consider only samples $\mathcal{S} \subset \mathcal{P}$ of n distinct units $u \in \mathcal{P}$.

a. *(1 mark)* Explain why, no matter what the attribute, when $n = N$ the sample error must be zero. (This is sometimes called Fisher consistency.)

   Solution:

   The sample error is calculated by:

   $$sampleError = a(S) - a(P_{study})$$

   When $n = N$, the sample S is same as the population P, therefore $a(S) = a(P_{study})$, which sum to give a total error of 0.

b. *(2 marks)* Suppose the attribute of interest is

   $$a_{min}(\mathcal{P}) = \min_{u \in \mathcal{P}}\ y(u)$$

   What is the largest possible sample error? And what sample $\mathcal{S}$ would produce it?.

   (If it helps, assume also that $y_i = y_j \iff i = j$ for all i and j in the population $\mathcal{P}$ – i.e. no tied y values.)

   Solution:

   The formula of sample error is $a(S) - a(P_{study})$. To get a large sample error, we have to make the $a(S)$ as big as possible. Based on the ordered variate values, we know that when $n = N$, we will have the largest $a(S)$. Therefore,

   $$sampleError = y_N - \min\ y(u)$$

   Substitute $min\ y(u)$ with y_1 in the formula, we get

   $$sampleError = y_N - y_1$$

c. *(4 marks)* Suppose the attribute of interest is now

   $$a_k(\mathcal{P}) = \frac{1}{N} \sum_{u \in \mathcal{P}} y_u^k$$

   for some $k > 0$ and let $\mathcal{C}$ denote the set of size $N_{\mathcal{C}}$ containing all possible samples $\mathcal{S}$ of n distinct units from $\mathcal{P}$.

   Prove that

   $$\frac{1}{N_{\mathcal{C}}} \sum_{\mathcal{S} \in \mathcal{C}} a_k(\mathcal{S}) = a_k(\mathcal{P}).$$

d. *(1 mark)* Given the result in part (c) is true, show that the *sampling bias* for these attributes is zero when the *sampling plan* is *simple random sampling* (without replacement).

Solution:

Given the formula:

$$SamplingBias = \overline{a}_C - a(P_study)$$

From part C, we know that $\frac{1}{N_{\mathcal{C}}}\sum_{\mathcal{S}\in\mathcal{C}} a_k(\mathcal{S}) = a_k(\mathcal{P})$. Substitute the results from part C to the sampling bias formula. We obtain:

$$SamplingBias = \frac{1}{N_{\mathcal{C}}}\sum_{\mathcal{S}\in\mathcal{C}} a_k(\mathcal{S}) - a_k(\mathcal{P})$$

. As a result, $SamplingBias = 0$