Papers on Anthropology XXX/2, 2021, pp. 9–13

IMPACT OF VOLUNTARY SAMPLING ON ESTIMATES

Ene-Margit Tiit

Statistics Estonia; University of Tartu, Tartu, Estonia

In sample-based surveys random sampling is the key issue. Only in the case of a random sample can the results obtained from the sample be generalized to the population of interest. Even in the case of any probabilistic sampling scheme with a complex design, the randomness of certain parts of the sample is crucial. Today, however, it is often difficult to obtain absolutely random samples, especially in the case of surveys where the people must participate as statistical units either answering to questionnaires or being measured. Nowadays, it is difficult to motivate randomly selected people to participate in a survey either to be measured or to respond to the survey questions.

One possible option, which is sometimes appropriate for researchers, is to use a voluntary sample. In most cases, however, this raises the issue of generalizability of the results: the voluntary sample is not representative of the population. For example, imagine that young women are invited to do an anthropometric survey. It is quite likely that women with a harmonious body size will participate in the survey with pleasure, whereas, for example, overweight girls will not participate, and there are no means to motivate them. Hence, the results of the sample survey cannot be generalized, the estimated values are shifted.

However, the problem of voluntary answering is much broader than in the case of voluntary samples as described above; it is something that researchers meet very often. The following two examples of a (partially) voluntary sample are quite common.

1. Sample with refusals. It almost always happens that the data of some objects (sampled persons) is missing for different reasons. Usually there are two ways to cope with the situation: either to use some imputation method to replace the missing values or to use a weighting scheme with the help of some basic characteristics. Both solutions are adequate if the non-response does not depend on the characteristics under investigation. However, very often this is not the case, as we can see from the example above.

https://doi.org/10.12697/poa.2021.30.2.01

2. In order to replace the non-response in the random sample, it has been supplemented by a voluntary sample, the distribution of which corresponds to the distribution of the non-response by some basic characteristics.

Both of these examples use the sample that is not random, and the parameter estimates and conclusions made using the samples may be skewed.

This article analyses situation 2 described above on the basis of theoretically constructed examples and demonstrates the dependence of the estimation bias on the ratio of parts of the sample.

In the case of sample surveys, sometimes a situation may arise that the researcher can use in addition to exactly designed (stratified random) sample $\boldsymbol{V}$ a set of additional measurement results (from the same population $\boldsymbol{P}$) for which the sampling rule is unknown. Also, there is no reason to assume that this dataset is a random sample. We call this additional dataset *voluntary sample* $\boldsymbol{F}$, the selection rule and distribution of which are unknown.

The researchers' question is – what to do with the so-called voluntary sample? In principle, three strategies can be implemented.

1. Only sample $\boldsymbol{V}$ is used.
2. Sample $\boldsymbol{F}$ will be used in the same way as sample $\boldsymbol{V}$, thereby increasing the size of the original sample.
3. Both samples are used but weighed differently.

In the first case, it is a waste of the data collected (accumulated), but the variance and bias of the estimate correspond exactly to the original plan. For the second and the third strategy options, the possible bias and variance of the estimates calculated need to be assessed in order to decide which strategy can give the optimal results. Obviously, this depends on both samples' volumes and distributions.

Assume that the aim of the study is to estimate the mean $\mathrm{E}X = m$ of variable X. For the sake of simplicity, we assume that the variance of X is the same in all samples, $\mathrm{D}(X) = s^2$.

Let the population size be $\#(\boldsymbol{P}) = N$ and let the sample sizes be $\#(\boldsymbol{V}) = n$ and $\#(\boldsymbol{F}) = \nu$, respectively.

In general, the conditional mean $a = \mathrm{E}\,(X|X \in F)$ of variable X in sample F differs from the population mean m of X.

Assume that the population consists of layers P_k, $k = 1, \ldots, K$ with sizes N_k, respectively, and the layers have conditional mean values according to $m_k = \mathrm{E}(X|X \in P_k)$, where

$$\mathrm{E}\sum_{k=1}^{K}\frac{N_k}{N}m_k = m.$$

Sample $\boldsymbol{V}$ consists of subsamples $\boldsymbol{V}_k$ belonging to strata $\boldsymbol{P}_k$, with sizes n_k, $k = 1, \ldots, K$, respectively.

For the first strategy, the sample is assumed to be layer-weighted, where the weights l_k^1 are determined by the standard rule

$$l_k^1 = \frac{N_k}{n_k}$$

and mean value $m(1)$ is unbiased.

In the second strategy, the numbers of voluntary observations in layer k are of size v_k, respectively, but the rule for selecting objects is unknown. For the sake of simplicity, we assume that volunteers are not included in sample $\boldsymbol{V}$ (this assumption is not restrictive, as it is always possible to use set $\boldsymbol{V} - \boldsymbol{F}$ instead $\boldsymbol{V}$ in calculations). Then, in layer k of the sample, the total number of observations is $n_k + v_k$. If the observations of both samples are considered to be equivalent, the weights

$$l_k^2 = \frac{N_k}{n_k + v_k}$$

should be used to calculate the estimates. However, since sample $\boldsymbol{F}$ is not random, the standard weighting does not warrant unbiased estimate and in the case of using the second strategy, the bias of the estimated mean $m(2)$ is $b(2) = \frac{v(a-m)}{n+v}$.

In the case of the third strategy, it is reasonable to assume that the points from sample $\boldsymbol{F}$ represent themselves only and, as it follows, their weights equal to one. In this case, the size of the population to be estimated is $N—v$ and the weights will have the following form:

$$\begin{cases} l_k^3 = \dfrac{N_k - v_k,}{n_k}. \text{ if } x \in \boldsymbol{V}, \\ \quad l_k^3 = 1, \text{if } x \in \boldsymbol{F}. \end{cases}$$

In general, the estimate of mean $m(3)$ is in this case not unbiased, but the estimated bias $b(3) = \frac{v(a-m)}{N}$ is rather small in the case of large N and moderate size v of sample ***F.***

When using different strategies, the variation of estimates also differs. A rough estimation is that, in the case of the second strategy, the variation of estimate decreases $\frac{n+v}{n}$ times, in the case of the third strategy, $-\frac{N}{N-v}$ times.

Example. The impact of a voluntary sample having a different mean in the case of different sample and population sizes.

Let the population size be N and there exist two samples – the designed sample of size n and the voluntary sample of size v. Assume that the aim of the survey is to estimate variable X having mean m and variance 1. The sample mean of the first sample is, by definition, m, but in the case of a voluntary sample, the sample mean a is different, assume that the difference is 1.

In Table 1 the biases and standard errors of estimates are calculated for strategies 1, 2 and 3.

Table 1. Using a combination of a random and a voluntary sample of different sizes for estimation.

				Bias			Standard error		
Population size *N*	Sample size *n*	Sample size *v*	Difference of means *a–m*	Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3
1000	100	10	1	0	0.09	0.01	0.1	0.095	0.099
1000	100	100	1	0	0.50	0.1	0.1	0.071	0.095
1000	100	500	1	0	0.83	0.5	0.1	0.041	0.071
10 000	2000	100	1	0	0.05	0.01	0.022	0.022	0.022
10 000	2000	1000	1	0	0.33	0.1	0.022	0.018	0.021
10 000	2000	5000	1	0	0.71	0.5	0.022	0.012	0.016

It is well-known that, when assessing the results, it is important to notice that from the bias of the estimated parameter follows, in general, an erroneous conclusion, while the bias or increase of the standard error means only somewhat wider confidence limits, but no error in conclusion.

From Table 1, it follows that, when considering the bias, then in all sample sizes the best result is gained using strategy 1 – that is, not to use voluntary sampling.

Comparing strategies 2 and 3, it is evident that in all cases strategy 3 gives better results than strategy 2, that means, when volunteers are used, they must

be weighted separately from the random sample. The largest bias occurs when the voluntary sample is used in the way as if it were a random sample.

The bias caused by the voluntary sample increases with the size of the voluntary sample compared to the random sample. Also, the bias increases when the sample size is large compared with the population size. The impact of the ratio of the two samples to standard error is not big.

From here, it follows that the researcher who plans to use a voluntary sample for completing a (too small) random sample should be rather cautious, as this step might cause biased results.

In practice, a situation, similar to the example, can arise when in a survey, besides the planned sample an additional sample will participate, using the internet. Usually this "easily motivated" additional sample represents a special subgroup of the population that might differ from the planned sample and cause biases in the results of the survey.

Address for correspondence:
Ene-Margit Tiit
Institute of Mathematics and Statistics
University of Tartu
Narva 18, 50409, Tartu, Estonia
E-mail: ene.tiit@ut.ee