**40th WEDC International Conference, Loughborough, UK, 2017**

LOCAL ACTION WITH INTERNATIONAL COOPERATION TO IMPROVE AND SUSTAIN WATER, SANITATION AND HYGIENE SERVICES

# Simplified sampling method for household-based surveys with reduced populations in the water and sanitation sector

*A. Pérez-Foguet & R. Giné-Garriga (Spain)*

**PAPER 2640**

*In order to make decisions efficiently and equitably, up-to-date information is required. In developing countries, with limited resources, such information should be provided by means of cost-effective methodologies, in which sampling issues are of primary importance. Different sampling strategies are currently in use. At local level with reduced populations, standard approaches prove expensive and time consuming. In this paper, we opt for simple linear piecewise approximations to calculate the sample size in terms of given precision, confidence level and population size. To support the applicability of the proposed approach by practitioners in the field, easy-to-use tables are elaborated. In terms of sampling, easy-to-follow practical guidelines for household selection and transect walk planning are also provided. The article presents six rural communities in Honduras as initial case study to illustrate the validity and applicability of the approach adopted herein for sampling design and sample size determination.*

## Introduction

Development strategies typically aim to ascertain whether or not a population in an area of influence meets certain welfare standards. To this end, decision-makers require accurate and up-to-date information to avoid decisions based on false assumptions. In the water and sanitation sector, this information is typically obtained through household-based surveys (Giné Garriga et al., 2013; Joint Monitoring Programme, 2006). With limited resources, one would like to take a representative sample of households and depending upon how many are covered by certain infrastructure, decide whether situation is adequate or whether additional efforts must be planned to improve the level of service delivered. The decision on the size of such sample is undoubtedly central to the sample design, as it affects both the precision and the cost of the survey (Bennett et al., 1991; United Nations Children's Fund, 2006).

In national surveys, where covering the overall study area would be practically impossible, various sampling methods are currently in use, including among others simple random sampling, stratified sampling and cluster sampling (Macro International Inc, 1996; United Nations Children's Fund, 2006). However, they present significant flaws if directly applied at lower administrative scales. Information needs to be highly disaggregated at the local scale, as the number of communities / villages is large (Grosh, 1997). And the population size in each administrative subunit is often reduced, since the number of households typically ranges from 20 to 500. With these figures, the direct application of the standards and guidelines commonly employed in large scale-surveys would produce too large samples, which in practice hinders the implementation of any local survey exercise.

Besides the sample size, a valid sampling procedure for the selection of households is also necessary to achieve reliable estimates. Mathematically speaking, the ideal procedure would be to have a list of all households in the community and choose a selection from the list at random. This exercise can be easily computed in any standard spread sheet. If such a list does not exist, it may be created by carrying out a quick census, e.g. by consulting community leaders, since total population will be reduced. Where this is not practicable, then a complete random exercise is not achievable. In order to ensure that the sample is as representative as possible, any method which achieves a random or near-random selection of households, preferably spread widely over the community, would be acceptable as long as it is clear and unambiguous, and does not give the field worker the opportunity to make personal choices which may introduce bias (Bennett et al., 1991). Experience suggests that a full list of households is rarely available, and both the elaboration of a census list and the random selection of a specific number of households - even with a spread sheet application - is not straightforward and time consuming.

Against this background, we present a simplified sampling method for household-based surveys where populations tend to be undersized, e.g. the common rural context in low-income countries. We determine sample sizes based on exact confidence intervals. To promote its practical application in the field, we develop an easy-to-use table to support practitioners when they need to decide on the appropriate sample size for a survey, i.e. when they need to strike the right balance between precision and cost. In terms of sampling, we then present easy-to-follow practical guidelines for a random household selection. We specifically opt for a transect walk across the community to ensure as wide a coverage as possible. In terms of implementation, advantage of this option is simplicity versus the "complexity" of undertaking a random selection. In the end, we show that although there are inherent limitations to the accuracy and precision of estimates, results may be used to classify and prioritize among groups of households. Such prioritization is valuable for local decision-making processes.

## Sample size determination

For populations that are large, usually in national surveys, the standard representative sample of size *n*, for a estimated proportion $p$, is given by (Cochran, 1977):

$$n = \left(\frac{z_{\alpha/2}}{d}\right)^2 p(1-p) \qquad (1)$$

where $\alpha$ is the confidence level, $z$ is a constant, which relates to the normally distributed estimator of the confidence level ($\alpha = 0.05$, $z_{\alpha/2}$ = 1.96; $\alpha = 0.1$, $z_{\alpha/2}$ = 1.64; $\alpha = 0.2$, $z_{\alpha/2}$ = 1.28), $d$ is the required precision on either side of the proportion, and a value of 0.5 is chosen for $p$ to maximize $p(1-p)$.

At the local level with reduced populations, however, the total population *N* is not significantly larger than *n*. As a result, previous Equation (1) cannot be applied. Alternatively, *n* may be computed from exact confidence limits for $p$, i.e. the Clooper-Pearson interval (Reiczigel, 2003). To obtain accurate confidence limits for reduced finite populations, we make use of the finite population correction ordinarily applied (Anderson and Burstein, 1968, 1967), but with minor fine-tuning (Burstein, 1975).

## Results

By adopting the previous approach for sample size determination, different sampling plans - shown in Table 1 - can be numerically computed in a standard spread-sheet. It is observed that sampling for a population size *N* lower than 10 may be to certain extent meaningless, since either the confidence level $\alpha$ or the precision *d* have to be significantly sacrificed. For instance, a sample size *n* of 7 in a population *N* of 8 would produce estimates within 33% (± 16.6%) of the true proportion with 80% confidence. And estimates with 90% and 95% confidence in a similar sampling design (*n*:*N* of 6:8) would fall within 25 percentage points of the true proportion. In contrast, a sample size of 12 from a population of size *N* lower than 100 guarantees estimates with 90% confidence within 25% points of precision. For higher precision, e.g. ± 12.5% or ± 10%, the sample size increases significantly up to 35 and 46 respectively; and with 95% confidence, the required size of the sample would be 15, 42 and 53. These figures can be compared with those obtained by applying standard Equation (1). It is worth noting that significant differences exist for large values of α, which indicates the limits of approximation by a normal distribution. In any case, the table becomes a useful instrument to select an optimal sample size on the basis of available resources, desired precision and the maximum permissible sampling error.

Results from Table 1 can be further exploited to prepare easy-to-use tables (see Table 2) that support practitioners in selecting and implementing in practice the most appropriate sampling plan. In the same way as in the previous Table 1, it is shown that there is little value in sampling for a population size N lower than 10, as almost all households must be targeted for sampling. It can also be easily inferred from the table that a sample size of 24 from a population of size 40 is needed to guarantee estimates with 90% confidence within 12.5% points of precision. Specifically, one would need to survey 6 out of 10 households of the community (6:10). At a practical level, this sampling plan could be applied through different sampling sequences, such as A-A-A-A-A-A-B-B-B-B - being "A" a surveyed household, and "B" a household not surveyed -, A-B-A-B-A-A-B-A-B-A, or A-A-B-A-B-A-A-B-A-B.

**Table 1. Sample size n for different values of N, α and d**

| N | $\alpha = 0.05$ | | | | $\alpha = 0.1$ | | | | $\alpha = 0.2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *d* < 1/10 | *d* < 1/8 | *d* < 1/6 | *d* < 1/4 | *d* < 1/10 | *d* < 1/8 | *d* < 1/6 | *d* < 1/4 | *d* < 1/10 | *d* < 1/8 | *d* < 1/6 | *d* < 1/4 |
| 8 | 8 | 8 | 8 | 6 | 8 | 8 | 8 | 6 | 8 | 8 | 7 | 5 |
| 10 | 10 | 10 | 9 | 7 | 10 | 10 | 9 | 7 | 10 | 9 | 8 | 6 |
| 15 | 14 | 14 | 12 | 9 | 14 | 13 | 11 | 8 | 14 | 12 | 10 | 7 |
| 20 | 18 | 17 | 15 | 10 | 18 | 16 | 13 | 9 | 17 | 15 | 12 | 7 |
| 25 | 22 | 20 | 17 | 11 | 21 | 19 | 15 | 9 | 19 | 16 | 13 | 8 |
| 30 | 25 | 23 | 18 | 12 | 24 | 21 | 16 | 10 | 21 | 18 | 13 | 8 |
| 40 | 31 | 27 | 21 | 13 | 29 | 24 | 18 | 10 | 25 | 20 | 14 | 8 |
| 50 | 36 | 31 | 23 | 13 | 33 | 27 | 20 | 11 | 28 | 22 | 15 | 8 |
| 75 | 46 | 37 | 27 | 14 | 40 | 32 | 22 | 12 | 32 | 25 | 16 | 9 |
| 100 | 53 | 42 | 29 | 15 | 46 | 35 | 23 | 12 | 35 | 26 | 17 | 9 |
| 150 | 64 | 48 | 31 | 16 | 53 | 39 | 25 | 12 | 39 | 28 | 18 | 9 |
| 250 | 75 | 54 | 34 | 16 | 60 | 42 | 26 | 12 | 43 | 30 | 18 | 9 |
| 500 | 87 | 60 | 36 | 17 | 67 | 46 | 27 | 13 | 46 | 31 | 19 | 9 |
| Eq. 1 | 97 | 62 | 35 | 16 | 68 | 44 | 25 | 11 | 42 | 27 | 15 | 7 |

**Table 2. Sampling plan for different values of N, α and d**

| *N* | $\alpha = 0.05$ | | | $\alpha = 0.1$ | | | $\alpha = 0.2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | *d* < 1/10 | *d* < 1/8 | *d* < 1/6 | *d* < 1/10 | *d* < 1/8 | *d* < 1/6 | *d* < 1/10 | *d* < 1/8 | *d* < 1/6 |
| 1-10 | All HHs | All HHs | All HHs | All HHs | All HHs | All HHs | All HHs | All HHs | All HHs |
| 11-20 | All HHs | 4:5 | 4:5 | All HHs | 4:5 | 4:5 | All HHs | 4:5 | 5:7 |
| 21-30 | 7:8 | 4:5 | 5:8 | 5:6 | 4:5 | 5:8 | 4:5 | 4:6 | 4:8 |
| 31-50 | 6:8 | 6:9 | 5:10 | 5:7 | 6:10 | 4:9 | 6:10 | 4:8 | 4:11 |
| 51-70 | 6:9 | 5:9 | 4:10 | 6:10 | 5:10 | 3:9 | 5:10 | 3:8 | 3:11 |
| 71-100 | 4:7 | 5:11 | 3:9 | 5:10 | 4:10 | 3:11 | 4:10 | 3:10 | 2:10 |

It has been verified that the decision about the choice of the sequence has an impact on achieved results. Indeed, a sampling plan *n*:*N* of 5:10 is preferred over a *n*:*N* of 1:2, given that the former promotes the randomness of the selection. In consequence, we have opted for considering groups of 5 to 11 households when defining the sampling plans. On the basis of this premise, it is remarkable that the enumerator will need to make two decisions when planning the transect walk: i) selection of the sampling sequence of households (as outlined above), and ii) selection of the first interviewed household. Both decisions may affect the results; and random or near-random selections should be therefore promoted.

Figure 1 compares different sampling plans presented in Table 2. The straight sloping line relate to a hypothetical census approach, where all households are included in the sample (n=N). The curve lines relate to different statistical significance levels when a sampling approach is adopted (each curve line is described by different values of d, α). For instance, a sample size n of 25 in a population N of 30 (i.e. 5 out of 6 households) should be included in a survey to produce estimates within 10% (± 5%) of the true proportion with 90% confidence (1/10, 0.1 line in the graph). By following this curve in the graph (i.e. same statistical precision: d=1/10, α=0.1), it is suggested that a sample size n of 36 would be required in a total population N of 50 (similar to the sampling plan *n*:*N* of 5:7 in Table 2). If precision of estimates can be lowered (e.g. $\alpha = 0.1$, d<1/8; 1/8, 0.1 line), then a sample size n of 30 would suffice to cover same population N of 50 (*n*:*N* of 6:10 in Table 2).

Figure 1 also shows for illustrative purposes the reference values that correspond to three different scenarios of effective workload in data collection (each scenario is described by a horizontal line). In the first scenario, a total sample n of 24 households are surveyed on a daily basis, e.g. a team of 2 enumerators, each one visiting 2 households per hour during 6 hours. In a population N of size 70, estimates with 90% confidence within 16.6% points of precision could be produced in one working day (1/6, 0.1 line in the graph). The second scenario is based on an overall sample n of 36 households per day, e.g. a team of 2 enumerators, each one visiting 3 households per hour during 6 hours. The increased productivity would provide a higher level of statistical significance and get estimates in the same population of size 70 within 25% (± 12.5%) of the true proportion with 90% confidence (1/8, 0.1 line). The last scenario (sample n of 48 households) could be implemented in one working day by a team of 2 enumerators, each one visiting 4 households per hour during 6 hours. This sampling approach would produce estimates with 90% confidence within 10% points of precision in a population N of 100 (1/10, 0.1 line).

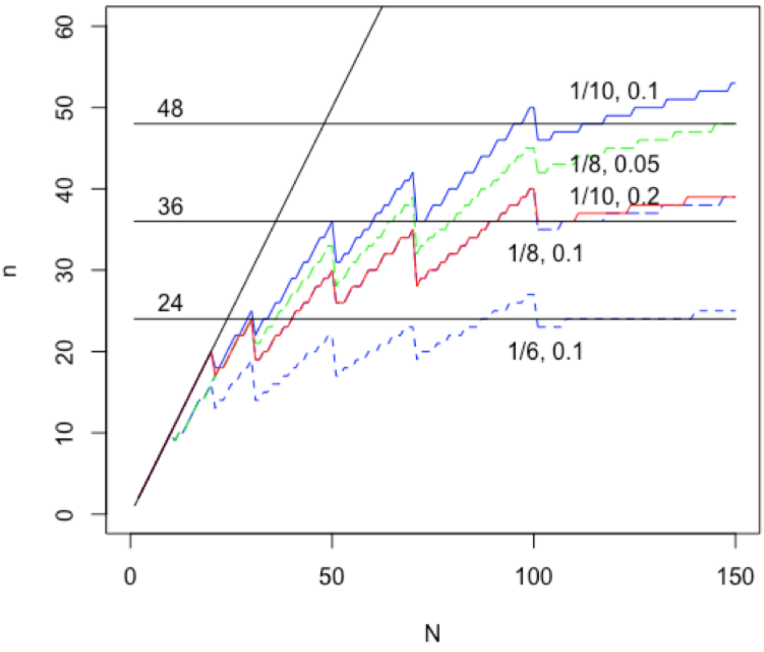

**Figure 1. Adjustment of the sample size n for different values of N, α and d**

## Discussion

For illustrative purposes, six rural villages with populations ranging from 11 to 42 households were surveyed in Honduras. In each community, we visited all households and carried out a quick questionnaire in relation to core sanitation and hygiene indicators. Specifically, three indicators were assessed: i) access to improved sanitation, as defined by the WHO/UNICEF Joint Monitoring Programme (Joint Monitoring Programme, 2008); ii) use of hand washing facility with soap and water at home; and iii) availability of in-house, clean and safe drinking-water storage containers.

The following table presents achieved results in the six visited villages (only results from one indicator are shown herein, i.e. handwashing device). First, we show the true percentage of HHs that verify the selected indicator from the whole sample in each village (i.e. the census). Then, estimated figures from two different sampling approaches are compared, namely the random sampling and a near random sampling through a transect walk. As regards to the estimates produced through a random selection of households, we present the average figure after performing a random selection of households up to 1,000 times. Second group of estimates are obtained by averaging results from simulating up to 1,000 times a near-random selection of households (transect walk), in which we have selected different sampling sequences and we have varied the first household. Remarkably, in this latter case, the sampling sequence of households in relation to the total population N will affect the sample size n, and the table shows the average sample size $n_{av}$ from all simulations.

**Table 3. Hand-washing device. Estimates based on the sampling plan: $\alpha$ = 0,1; d = 1/8 or d = 1/6 (depending on the total population N of each village).**

| Sampling | | Village A | Village B | Village C | Village D | Village E | Village F |
|---|---|---|---|---|---|---|---|
| **Census** | **N / y** | 44 / 24 | 38 / 29 | 38 / 30 | 42 / 33 | 13 / 5 | 11 / 2 |
| | **p** | **0,545** | **0,763** | **0,789** | **0,786** | **0,385** | **0,182** |
| **Required Precision** | **d** | 0,125 | 0,125 | 0,125 | 0,125 | 0,167 | 0,167 |
| **Random Approach** | **n / n:N** | 25 / 6:10 | 24 / 6:10 | 24 / 6:10 | 25 / 6:10 | 10 / 4:5 | 9 / 4:5 |
| | **max \|pu' - pl'\| /2** | 0,124 | 0,115 | 0,112 | 0,116 | 0,164 | 0,136 |
| | **$p'_{av}$** | **0,543** | **0,765** | **0,789** | **0,784** | **0,382** | **0,178** |
| | **$pl'_{av}$** | 0,418 | 0,650 | 0,675 | 0,668 | 0,236 | 0,081 |
| | **$pu'_{av}$** | 0,663 | 0,853 | 0,872 | 0,869 | 0,549 | 0,334 |
| | **max \|p' – p\|** | 0,215 | 0,195 | 0,169 | 0,174 | 0,185 | 0,182 |
| | **Count_no_IC** | 0,074 | 0,054 | 0,027 | 0,067 | 0,034 | 0,021 |
| **Transect Walk** | **$n_{av}$** | 26,401 | 22,799 | 22,803 | 25,197 | 10,446 | 8,877 |
| | **max \|pu' - pl'\| /2** | 0,130 | 0,123 | 0,127 | 0,120 | 0,159 | 0,173 |
| | **$p'_{av}$** | **0,545** | **0,763** | **0,790** | **0,782** | **0,387** | **0,187** |
| | **$pl'_{av}$** | 0,427 | 0,639 | 0,668 | 0,667 | 0,251 | 0,085 |
| | **$pu'_{av}$** | 0,658 | 0,857 | 0,877 | 0,867 | 0,541 | 0,348 |
| | **max \|p' – p\|** | 0,161 | 0,106 | 0,167 | 0,146 | 0,115 | 0,182 |
| | **Count_no_IC** | 0,042 | 0,000 | 0,053 | 0,046 | 0,015 | 0,047 |

Notes: y: number of surveyed HHs that verify the selected indicator; p: proportion of HHs that verify the selected indicator (y/N); pl': Estimated lower confidence limit; pu': Estimated upper confidence limit; $p'_{av}$: Average p value after performing a random / near random selection of households up to 1,000 times; $pl'_{av}$: Average of estimated lower confidence limit after performing a random / near random selection of households up to 1,000 times; $pu'_{av}$: Average of estimated upper confidence limit after performing a random / near random selection of households up to 1,000 times; Count_no_IC: Number of cases where the proportion p' does not fall within the expected confidence intervals.

As one would expect, it is observed that the results achieved from a census approach are consistent with those achieved from a sampling approach in all surveyed villages. Specifically, the differences obtained are negligible in percentage values (see p' in comparison to p). It is shown that the largest difference between p' (estimated) and p (true proportion) - see max$|$p' – p$|$ - is not considerably higher than d. In addition, the real proportions p' are in the great majority of cases within the confidence interval of the estimate (see the parameter Count_no_IC), and in all villages, the % of cases where this does not happen is lower than the level of confidence (i.e. 10%). In all, it might be said that reduced sample sizes in small populations are adequate to produce estimates that are sufficiently precise to support local decision-making. In addition, no significant differences are observed in the method employed for household selection, i.e. the pure random selection and the transect walk.

## Conclusions

In an era of increasing decentralization of basic services, the need for reliable performance data at the local level is emerging. In particular, decision-makers need to identify the neediest areas to allot resources efficiently, equitably and transparently. Many field data collection methodologies with different sampling strategies have been developed in recent years, but when applied in decentralized contexts with reduced populations, they present significant shortcomings. There is the need to balance precision of survey data against survey costs, and of primary importance is the question of how large should the sample be in order to produce precise confidence intervals for the estimates obtained.

The aim of this study is to adopt a simplified approach to sample size determination for local surveys. Specifically, it defines the sampling plan required to assess household-related issues in small populations; i.e. it helps select the minimum sample size on the basis of desired precision and the maximum permissible sampling error. From a practitioner viewpoint, it should help select the sampling plan that best confronts the dilemma between precision and cost. In addition, the article suggests an easy-to-adopt approach for the selection of households through a transect walk in the community.

Based on one case study from Honduras, the results indicate that reduced sample sizes may be valid to produce estimates with sufficient precision to be used in decision-making processes.

### Acknowledgements

The authors are grateful to A. Jiménez (Stockholm International Water Institute) and O. Flores for their useful comments and suggestions. Further thanks go to MSc students D. Requejo and H. Grau, as well as to Honduran organizations, for their support to undertake the field work. Financial support from the Catalan Government (Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), project ref: 2014SGR1545:2014-2016) and the Centre de Cooperació per al Desenvolupament Universitat Politècnica de Catalunya (grant ref 2015-O034) is gratefully acknowledged.

### References

Anderson, T.W., Burstein, H., 1968. Approximating the Lower Binomial Confidence Limit. J. Am. Stat. Assoc. 63, 1413–1415.

Anderson, T.W., Burstein, H., 1967. Approximating the Upper Binomial Confidence Limit. J. Am. Stat. Assoc. 62, 857–861.

Bennett, S., Woods, T., Liyanage, W.M., Smith, D.L., 1991. A simplified general method for cluster-sample surveys of health in developing countries. World Heal. Stat Q 44, 98–106.

Burstein, H., 1975. Finite Population Correction for Binomial Confidence Limits . J. Am. Stat. Assoc. 70, 67–69. doi:10.2307/2285378

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. John Wiley and Sons, New York.

Giné Garriga, R., Jiménez, A., Pérez Foguet, A., 2013. Water-sanitation-hygiene mapping: An improved approach for data collection at local level. Sci. Total Environ. 463–464, 700–711.

Grosh, M.E., 1997. The policymaking uses of multitopic household survey data: A primer. World Bank Res. Obs. 12, 137–160.

Joint Monitoring Programme, 2008. Progress on Drinking Water and Sanitation: Special Focus on Sanitation, Joint Monitoring Programme for Water Supply and Sanitation. WHO / UNICEF, Geneva / New York.

Joint Monitoring Programme, 2006. Core questions on drinking-water and sanitation for household surveys. WHO / UNICEF, Geneva / New York.

Macro International Inc, 1996. Sampling Manual, DHS-III. ed. Macro International Inc, Calverton, Maryland.

Reiczigel, J., 2003. Confidence intervals for the binomial parameter: some new considerations. Stat. Med. 22, 611–621. doi:10.1002/sim.1320

United Nations Children's Fund, 2006. Multiple Indicator Cluster Survey Manual 2005. UNICEF, Division of Policy and Planning, New York.

## Contact details

A. Pérez Foguet & R. Giné Garriga
Universitat Politècnica de Catalunya (UPC), Spain
Email: agusti.perez@upc.edu / ricard.gine@upc.edu
www: http://www.engsc-gdev.cat/