RECOMMENDATIONS ON SIMPLIFIED STATISTICAL METHODS USED FOR HUMANITARIAN PROJECTS

Jean Lapegue – Senior Advisor WASH ACF-France

1 - Introduction to robust statistical and sampling methods

Humanitarian projects are increasingly subjected to produce proofs for the needs or the impact of their activities. Furthermore, it is increasingly common to publish these data within the context of scientific publications or presentations at scientific conferences

Robust statistical procedures which still remain simple and practically applicable in the field, are essential for humanitarian actors such as ACF, thus complementing the possibility of developing its action projects whose quality and technical relevance are recognised.

On the other hand, conducting a robust survey or a real statistical analysis is not necessarily more complex or does not take much time than trying an approximate method. A statistical analysis or a non-robust survey renders the results useless, impossible to publish, export and potentially discredits projects and organization.

In order to conduct robust surveys, particular care should be taken when selecting the sample of a research study . This step is crucial in order to be able to interpret the results of the research based on a population sampling that is representative of total population.

First of all, we will select the most appropriate sampling technique by taking into account the following:

a) Main objectives of the study carried out (which will determine the degree of accuracy),

b) The characteristics of the study population (size, differentiated groups)

c) Field constraints.

Beyond this notion of representativeness, sampling principle implies that all individuals within a concerned population must have, at best, the same probability of taking part of the selected sample.

2 - Definitions

Population and individual: P

Comprehensive set of individuals in the area covered by the survey (there may be individuals, families, homes, infrastructure, etc.). An element of this set is called individual.

Sample N

Group of individuals who will be interviewed. The sample N is supposed to be representative of the surveyed P population.

Character and modalities

The common feature that is being studied is called character.The values taken by the character are also called modalities. The difference of the extreme values of the character is called range.

Effective (ni) and frequency (fi)

The number of individuals ni of a survey is called effective.The total number of individuals in the surveyed population is called total number (N).

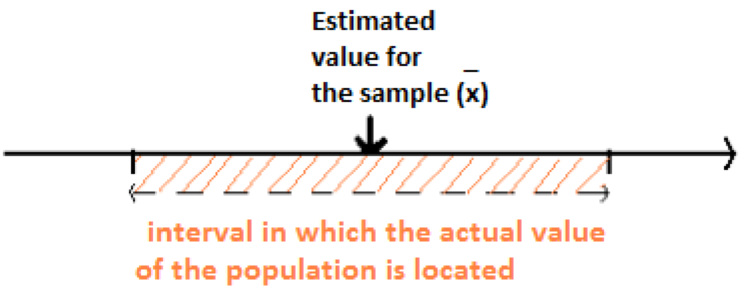
The ratio fi = ni / N is called frequency. fi is always a number between 0 and 1 (expressed in%). The sum of the numbers fi is always equal to 1.

Degree of Precision d[1]

The degree of precision estimates confidence interval in which the value of the actual population is located. The latter is expressed in percentage points (eg: + / - 5%).

(Example: if the rate of diarrhea in children in the sample is 20% with an accuracy degree of + / - 5%, then we can say that the rate is between 15% and 25% for the total child population).

We'll choose 5% in the case of epidemiological studies (biological water quality, diarrhea rates, etc..) Or 10% in all other cases (CAP survey, health survey, etc..).


Estimated value for the sample ($\bar{x}$)
interval in which the actual value of the population is located

Confidence interval = [X-d; X+d

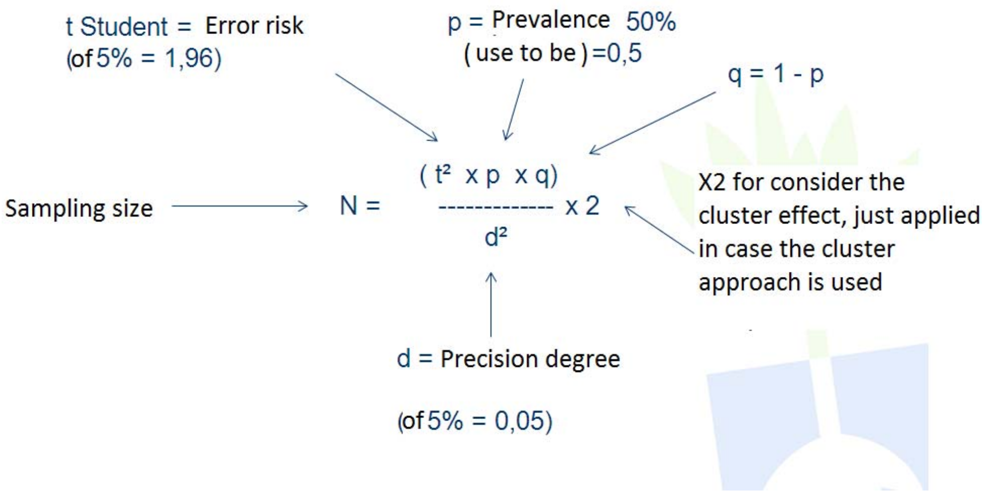
[1] For example, for a cluster method with a 10% degree of precision accuracy, and beyond a homogeneous population of 1500 individuals, the sample size is not based on population.

Representativeness:

Capacity of the sample to represent the population. The sample size only contributes to a certain extent to the representativeness of the population [1].

3 - Classical Formula for sample size calculation

You need to know how to determine this formula, but we prefer to use the table (section 6) to estimate the desired sample size, depending on the method chosen.



4 - Different standard methods of sampling

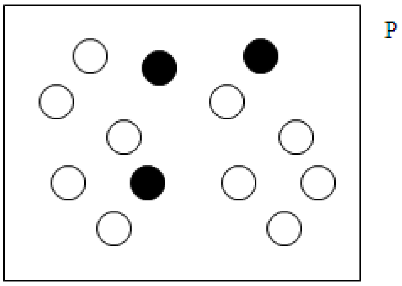
We only present here the two methods that should be used and acknowledged in priority[2], which meet the majority of practical cases of the field.

a) simple random method, use only if you have the nominal list of all individuals, which is rare.

b) The method by cluster is strongly recommended in all cases

● Surveyed individual

○ Unsurveyed individual



a) <u>Simple Random</u>

Method only possible if we have a list of individuals (eg family names of a refugee camp, hospital patients, etc.). The choice of individuals in the sample must be randomly (not by pulling names at random but using eg Excel statistical function **RAND.BETWEEN(x,y)** in your Excel spreadsheet.

[2] We will not voluntarily discuss about other random methods here (weighted sampling, proportional stratified random sampling) or non-random (purposive sampling, systematic and cumulative), or even mixed methods combining the methods by cluster and random methods

b) Cluster

This is the method that should be used primarily in the context of statistical surveys related to humanitarian projects. This method is recommended by WHO and UNICEF (nutrition surveys). It is simple, fast and robust. The population is divided into 30 "clusters" each represented by 5-7 individuals [3] (for a degree of accuracy of 10% which applies to CAP surveys or health survey) or 15 to 26 individuals (for a degree of precision of 5% which is well suited for epidemiological studies, biological quality, etc.). Therefore, 210 or 780 interviews depending on the level chosen.

Note: if you want to compare two groups of the same population (eg people who have benefited from a ACF project and those who have not), the chosen statistical method will have to be applied to each of the two groups. With a cluster method, and a degree of accuracy of 10%, this brings the number of individuals interviewed to 420 (2 x 210).

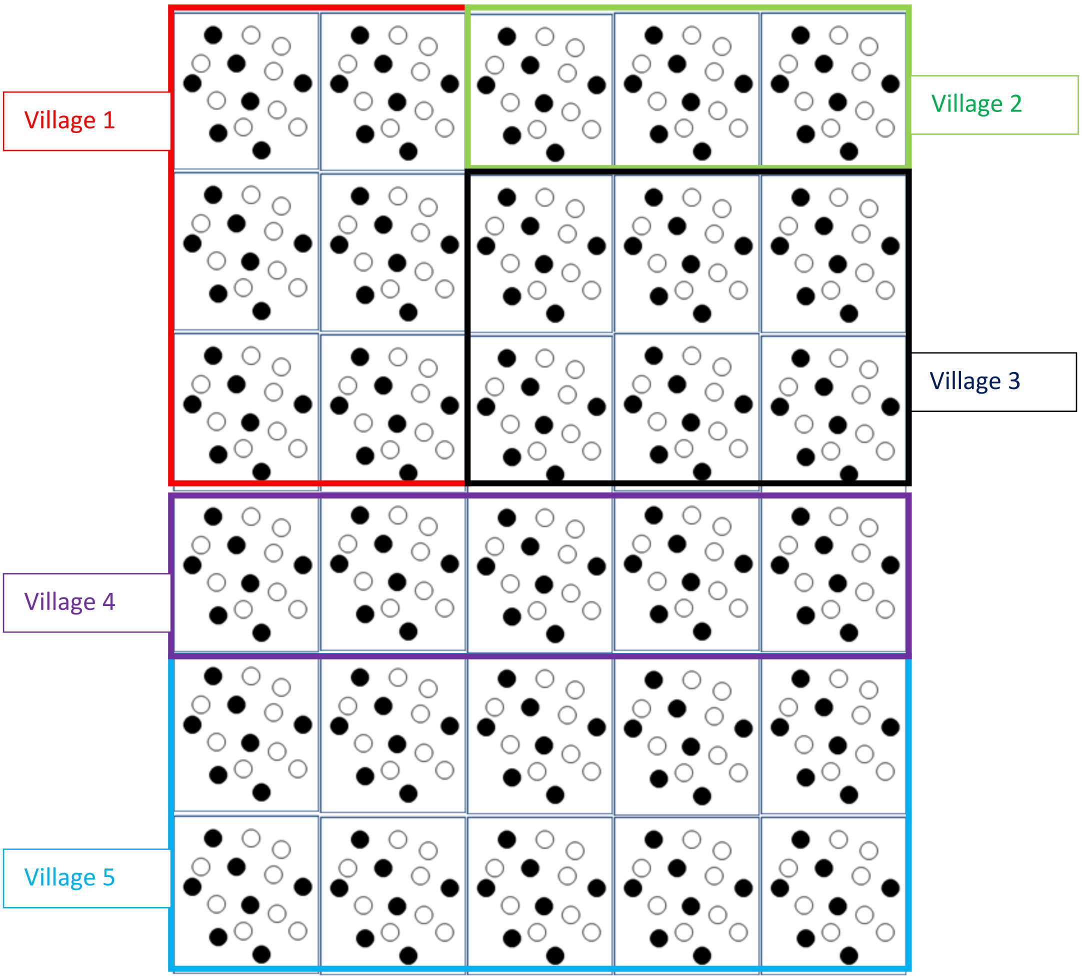


Figure: Distribution of 30 clusters each containing seven interviewed individuals, representation of the total population of the area (five villages)

[3] See Table section 6 for the number of individuals interviewed by cluster, which depends on the size of the population

5 - Practical Exercise (cluster approach)

ACF stars a WASH project in a population of 158,070 people, scattered among four large villages of Northwest CAR. Calculate the number of interviews required in every village to have a good representation of the results of a CAP survey?[4]

Questions posed by the interviewer *(In blue, proposed answer)*

Q1-Which method: random or cluster? Cluster (as there's no nominal list of households)

Q2-How accurate? 10% (no epidemiological dimension within a CAP)

Q3-What is the sample size N? 210 individuals (families) will be interviewed

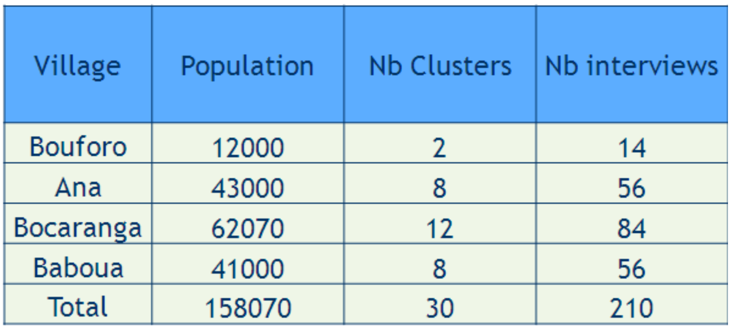
Q4-How many clusters? 30 (classical method UNICEF-WHO)

Q5-How many interviewed households per cluster? 7 (UNICEF-WHO 10% method)

Q7-How many interviews will take place in town?

a) Each of the 30 clusters will represent 158070/30 = 5269 people

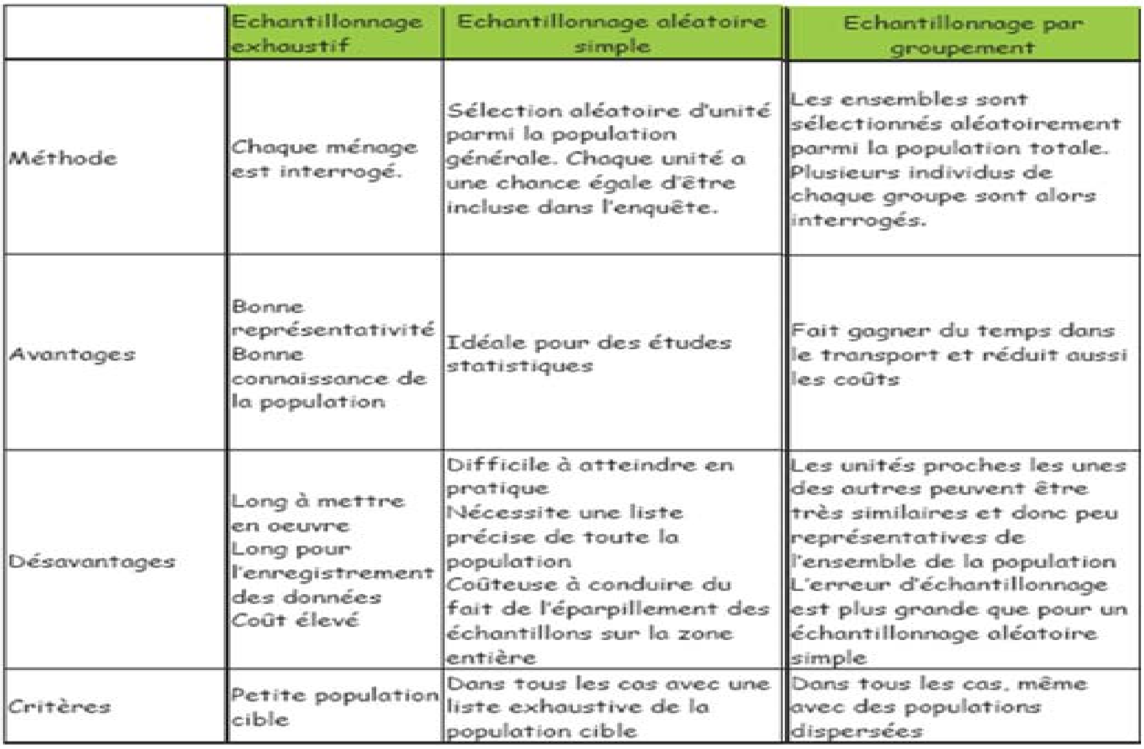
b) The distribution of clusters is intuitively done like this:



Village	Population	Nb Clusters	Nb interviews
Bouforo	12000	2	14
Ana	43000	8	56
Bocaranga	62070	12	84
Baboua	41000	8	56
Total	158070	30	210

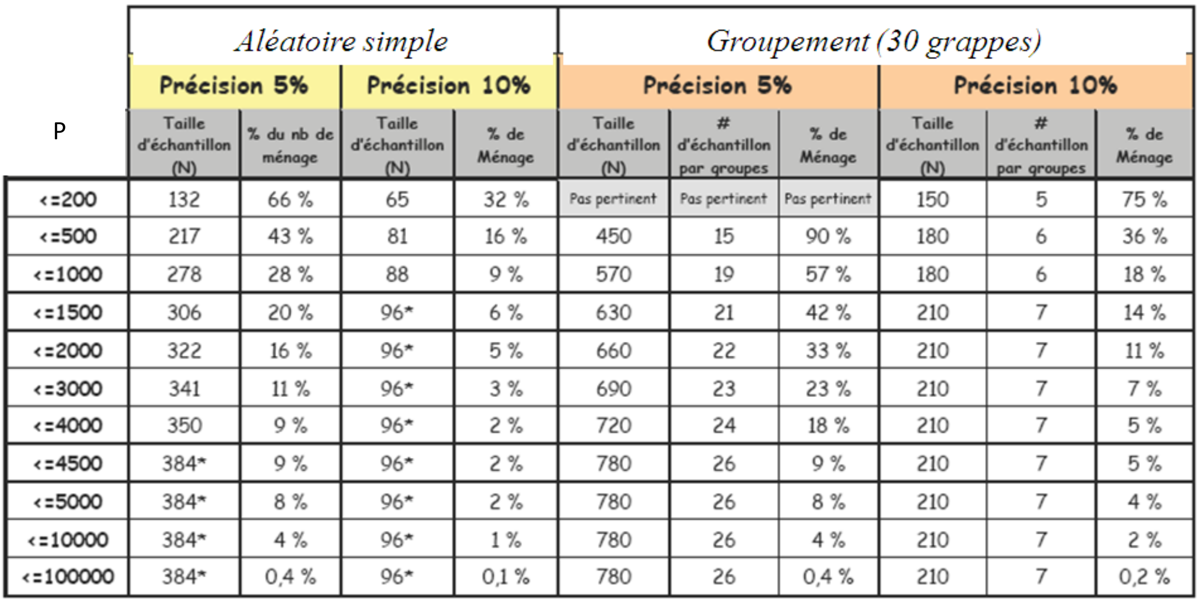
[4] We do not have details of the nominal list of households in this simulation

6 - Criteria for selecting sampling method



	Echantillonnage exhaustif	Echantillonnage aléatoire simple	Echantillonnage par groupement
Méthode	Chaque ménage est interrogé.	Sélection aléatoire d'unité parmi la population générale. Chaque unité a une chance égale d'être incluse dans l'enquête.	Les ensembles sont sélectionnés aléatoirement parmi la population totale. Plusieurs individus de chaque groupe sont alors interrogés.
Avantages	Bonne représentativité Bonne connaissance de la population	Idéale pour des études statistiques	Fait gagner du temps dans le transport et réduit aussi les coûts
Désavantages	Long à mettre en oeuvre Long pour l'enregistrement des données Coût élevé	Difficile à atteindre en pratique Nécessite une liste précise de toute la population Coûteuse à conduire du fait de l'éparpillement des échantillons sur la zone entière	Les unités proches les unes des autres peuvent être très similaires et donc peu représentatives de l'ensemble de la population L'erreur d'échantillonnage est plus grande que pour un échantillonnage aléatoire simple
Critères	Petite population cible	Dans tous les cas avec une liste exhaustive de la population cible	Dans tous les cas, même avec des populations dispersées

7 - Sampling: table for simple random methods, cluster and stratified



	Aléatoire simple				*Groupement (30 grappes)*					
	Précision 5%		Précision 10%		Précision 5%			Précision 10%		
P	Taille d'échantillon (N)	% du nb de ménage	Taille d'échantillon (N)	% de Ménage	Taille d'échantillon (N)	# d'échantillon par groupes	% de Ménage	Taille d'échantillon (N)	# d'échantillon par groupes	% de Ménage
<=200	132	66 %	65	32 %	Pas pertinent	Pas pertinent	Pas pertinent	150	5	75 %
<=500	217	43 %	81	16 %	450	15	90 %	180	6	36 %
<=1000	278	28 %	88	9 %	570	19	57 %	180	6	18 %
<=1500	306	20 %	96*	6 %	630	21	42 %	210	7	14 %
<=2000	322	16 %	96*	5 %	660	22	33 %	210	7	11 %
<=3000	341	11 %	96*	3 %	690	23	23 %	210	7	7 %
<=4000	350	9 %	96*	2 %	720	24	18 %	210	7	5 %
<=4500	384*	9 %	96*	2 %	780	26	9 %	210	7	5 %
<=5000	384*	8 %	96*	2 %	780	26	8 %	210	7	4 %
<=10000	384*	4 %	96*	1 %	780	26	4 %	210	7	2 %
<=100000	384*	0,4 %	96*	0,1 %	780	26	0,4 %	210	7	0,2 %

8 - Results presentation, test group, ethical strategies for measuring impact

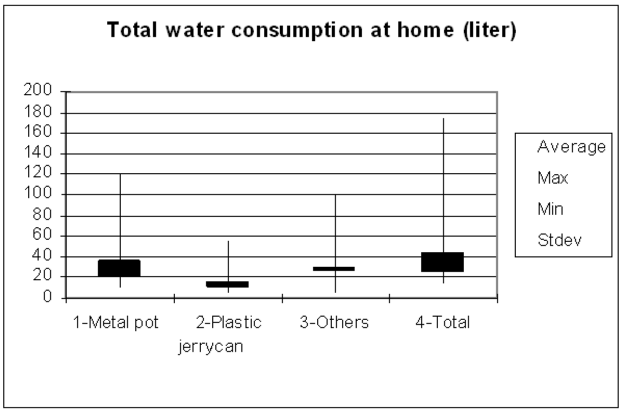
In presenting the results of the statistical survey:

a) Always explain the choice of the methodology and justification (accuracy, method chosen, etc.)

b) When working with two groups (a project group which benefits from ACF projects and a 'control' group which does not benefit from ACF projects) group must maintain an ethical approach must be preserved (for instance, the same group will be tested before and after a project).

c) When you want to compare two groups (case of heterogeneous area) each tested groups will have to benefit from a robust method, otherwise , it will not be compared.

d) The results of the survey must be systematically presented along with their confidence intervals (standard deviation 'in English). This **STDDEVIATION(x,y)** function is easy to find in Excel (see Figure below).



· Total water consumption by day

e) When comparing two surveys (e.g. pre and post CAPs project) it is recommended to interview the same individuals.