

RHRC CONSORTIUM MONITORING AND EVALUATION TOOLKIT
SURVEY SAMPLING TECHNIQUE EXAMPLE

**Instructions for
Probability Proportional to Size Sampling Technique**

*Prepared by Therese McGinn
Heilbrunn Department of Population and Family Health
Mailman School of Public Health, Columbia University*

Probability proportional to size (PPS) is a sampling technique for use with surveys or mini-surveys in which the probability of selecting a sampling unit (e.g., village, zone, district, health center) is proportional to the size of its population. It gives a probability (i.e., random, representative) sample.

It is most useful when the sampling units vary considerably in size because it assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice versa. This method also facilitates planning for field work because a pre-determined number of respondents is interviewed in each unit selected, and staff can be allocated accordingly.

Steps in Applying Probability Proportional to Size Sampling

1. List all villages in the project area (Column A in the following example) and their populations (Column B). You can use the total population or the population of the group you are sampling, for example, married women aged 15-44 or men 15-60. In a pinch, the number of households can be used.
2. Calculate the running cumulative population (Column C). The last number in this column is the total population of the project area. In the sample, the total population is 17,619.
3. Determine the number of sites which will be visited and the total sample size desired. For a mini-survey, from which only basic frequencies are desired, expect to visit 10-15 villages for a total sample size of 100-200. For a full scale survey, the sample size will be determined by the level and type of analysis planned; you should probably expect to visit 15-30 sites, although the final number will be determined by the complexity of the area and the purpose of the study.

In this example, we will visit 10 sites to conduct a mini-survey with a desired sample size of 150 women aged 15-44. Thus, 15 women will be interviewed in each of the 10 sites selected.

4. Divide the total population of the project area (17,619, the final figure in Column C) by 10, the number of sites we decide to visit. The result, 1,762, is called *the Sampling Interval (SI)*.
5. Choose a number between 1 and the SI at random. This is the *Random Start (RS)*. In this sample, the RS is 1321.

6. Calculate the following series: RS; RS + SI; RS + 2SI; RS + 3SI; RS + 4SI; RS + 5SI; RS + 6SI; RS + 7SI; RS + 8SI; RS + 9SI.

Example: RS + 2SI is to be calculated as 2 times the sampling interval added to the random start. In this case, $1321 + 2(1762) = 4845$.

7. Each of these 10 numbers corresponds to a site on the list of villages. The villages selected are those for which Column C, the cumulative population, contains the numbers in the series we calculated.

For example, the first number in the series, 1,321, is contained in village 3, which holds numbers 788 to 1,819. The second number in the series (3,083) is contained in village 6, which holds numbers 2,943 to 3,294.

Continuing in this manner, the desired number of sites will be selected. In this example, the selected villages are numbers 3, 6, 9, 11, 15, 18, 21, 22, 25 and 29 (Column D).

8. As planned, 15 interviews will be conducted in each of the 10 villages selected. Selection of respondents within the village should also be done randomly, preferably from a list of eligible names or a map of households. If these are not available, estimate the number of households in the village from the population figures, then divide that number by 15, the number of respondents desired. This is the interval, n . Starting from a random household, count every n th household and interview all eligible respondents in that household. For example, in village 3 there are 1,032 people. If other information suggests that an average of 6 people make up a household, then we estimate that there are 172 household in the village ($1032/6=172$). To get 15 respondents, we need to sample every 11th household ($172/15=11.5$).

Notes

- It can happen that a very large village contains more than one of the series of numbers. In this case, that village counts as two sites and twice the allocated number of interviews should be conducted there.
- Once a household has been selected into the sample, *all* eligible respondents in the household should be interviewed. For example, if the survey seeks women aged 15-44 and one household contains a mother, 42, and her daughter, 19, then *both* should be interviewed. The reason for this is that if only one were interviewed, then a woman living in a household with another woman would have a lower chance (only 1 in 2) of getting into the sample than a woman living with no other eligibles, whose chance is 1 in 1. The former group would be systematically deselected from the sample. Such a situation could arise with polygamous women, mothers and daughters or live-in servants. Such systematic bias could affect the data.
- It is better to exceed the sample size than not to reach it. In the above case, 15 households may result in more than 15 interviews. Depending on the frequency of the situation, you can leave the plan as is and end up with more interviews than you really need or you can compensate for the multiple cases by sampling fewer households.

**Example: Drawing a Sample Using
Probability Proportional to Size Sampling Technique**

<u>Column A</u>	<u>Column B</u>	<u>Column C</u>	<u>Column D</u>
Village	Village Population	Cumulative Population	Series Numbers/ Selected Sites
1	542	542	
2	245	787	
3	1032	1819	1321
4	867	2686	
5	256	2942	
6	352	3294	3083
7	835	4129	
8	645	4774	
9	427	5201	4845
10	312	5513	
11	1342	6855	6607
12	390	7245	
13	604	7849	
14	465	8314	
15	897	9211	8369
16	476	9687	
17	365	10052	
18	967	11019	10131
19	533	11552	
20	215	11767	
21	1590	13357	11893
22	423	13780	13655
23	645	14425	
24	867	15292	
25	423	15715	15417
26	197	15912	
27	586	16498	
28	365	16863	
29	756	17619	17179

$$\begin{aligned} \text{Sampling Interval (SI)} &= \text{Cumulative population} / \text{Number of sites} \\ &= 17619 / 10 \\ &= 1762 \end{aligned}$$

$$\text{Random Start (RS)} = 1321$$

Series numbers	RS	1321	RS+5SI	10131
	RS+SI	3083	RS+6SI	11893
	RS+2SI	4845	RS+7SI	13655
	RS+3SI	6607	RS+8SI	15417
	RS+4SI	8369	RS+9SI	17179

Selected sites are villages 3, 6, 9, 11, 15, 18, 21, 22, 25, 29