Steps in applying Probability Proportional to Size (PPS)

and calculating Basic Probability Weights

First stage:	PPS sampling → larger clusters have bigger probability of being sampled
Second stage:	Sampling exactly the same number of individuals per cluster → individuals in large clusters have smaller probability of being sampled
Overall:	Second stage compensates first stage, so that each individual in the population has the same probability of being sampled

1. Calculate the sample size for each strata.
2. Separate population data into strata. The following steps will have to be applied for each strata.
3. List the primary sampling units (Column A) and their population sizes (Column B). Each cluster has its own *Cluster Population Size (a)*.
4. Calculate the cumulative sum of the population sizes (Column C). The *Total Population (b)* will be the last figure in Column C.
5. Determine the *Number of Clusters (d)* that will be sampled in each strata.
6. Determine the *Number of Individuals to be sampled from each cluster (c)*. In order to ensure that all individuals in the population have the same probability of selection irrespective of the size of their cluster, the same number of individuals has to be sampled from each cluster.
7. Divide the total population by the number of clusters to be sampled, to get the *Sampling Interval (SI)*.

8. Choose a random number between 1 and the *SI*. This is the *Random Start (RS)*. The first cluster to be sampled contains this cumulative population (Column C). [Excel command =rand()*SI]
9. Calculate the following series: *RS; RS + SI; RS + 2SI; …. RS+(d-1)*SI.*
10. The clusters selected are those for which the cumulative population (Column C) contains one of the serial numbers calculated in item 8. Depending on the population size of the cluster, it is possible that big clusters will be sampled more than once. Mark the sampled clusters in another column (Column D).
11. Calculate for each of the sampled clusters the *Probability of Each Cluster Being Sampled (Prob 1)* (Column E).

    Prob 1= (a x d) ÷ b

    a= Cluster population

    b= Total Population

    d= Number of Clusters

12. Calculate for each of the sampled clusters the *Probability of each individual being sampled in each cluster (Prob 2)* (Column G).

    Prob 2= c / a

    a= Cluster population

    c= Number of individuals to be sampled in each cluster

13. Calculate the overall basic weight of an individual being sampled in the population. The basic weight is the inverse of the probability of selection.

    BW=1/(prob 1 * prob 2)

Example:
Population 20000 in 30 clusters.
Sample 3000 from 10 clusters using PPS.
Calculate Prob. 1 = probability of selection for each sampled cluster,
Calculate Prob. 2 = probability of selection for each individual in each of the sampled clusters,
Calculate the overall weight = inverse of the probability of each individual being sampled in the population

A	B	C	D	E	F	G	H
Cluster	**Size (a)**	**Cumulative sum**	**Clusters sampled**	**Prob 1**	**Individuals per cluster (c)**	**Prob 2**	**Overall weight**
1	1028	1028	905	51%	300	29%	6.7
2	555	1583					
3	390	1973					
4	1309	3282	2905	65%	300	23%	6.7
5	698	3980					
6	907	4887					
7	432	5319	4905	22%	300	69%	6.7
8	897	6216					
9	677	6893					
10	501	7394	6905	25%	300	60%	6.7
11	867	8261					
12	867	9128	8905	43%	300	35%	6.7
13	1002	10130					
14	1094	11224	10905	55%	300	27%	6.7
15	668	11892					
16	500	12392					
17	835	13227	12905	42%	300	36%	6.7
18	396	13623					
19	630	14253					
20	483	14736					
21	319	15055	14905	16%	300	94%	6.7
22	569	15624					
23	987	16611					
24	598	17209	16905	30%	300	50%	6.7
25	375	17584					
26	387	17971					
27	465	18436					
28	751	19187	18905	38%	300	40%	6.7
29	365	19552					
30	448	20000 (b)					

Number of clusters (d) = 10

Sampling interval (SI) = Cumulative population (B) / Number clusters (D) = 20000/10 = 2000

Random Start (RS) = 905

Series numbers

1	RS=	905	6	RS+(5*SI)=	10905
2	RS+(1*SI)=	2905	7	RS+(6*SI)=	12905
3	RS+(2*SI)=	4905	8	RS+(7*SI)=	14905
4	RS+(3*SI)=	6905	9	RS+(8*SI)=	16905
5	RS+(4*SI)=	8905	10	RS+(9*SI)=	18905

Probability 1 = (a*d) / b Probability of selection for each sampled cluster

Probability 2 = c / a Probability of selection for each individual in each of the sampled clusters

Overall weight = 1 / (Prob1 * Prob2) Inverse of the probability of each individual being sampled in the population

a= number individuals in each cluster
b=sum individuals in all clusters
c=number individuals sampled per cluster
d=number sampled clusters

Definitions

The sampling frame is the list of ultimate sampling units, which may be people, households, organizations, or other units of analysis.

Random sampling is data collection in which every person in the population has a chance of being selected which is known in advance. Normally this is an equal chance of being selected. Random samples are always strongly preferred, as only random samples permit statistical inference.

Probability proportion to size is a sampling procedure under which the probability of a unit being selected is proportional to the size of the ultimate unit, giving larger clusters a greater probability of selection and smaller clusters a lower probability. In order to ensure that all units (ex. individuals) in the population have the same probability of selection irrespective of the size of their cluster, each of the hierarchical levels prior to the ultimate level has to be sampled according to the size of ultimate units it contains, but the same number of units has to be sampled from each cluster at the last hierarchical level. This method also facilitates planning for field work because a pre-determined number of individuals is interviewed in each unit selected, and staff can be allocated accordingly

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 verse.

The design effect (D) is a coefficient which reflects how sampling design affects the computation of significance levels compared to simple random sampling (discussed below). A design effect coefficient of 1.0 means the sampling design is equivalent to simple random sampling. A design effect greater than 1.0 means the sampling design reduces precision of estimate compared to simple random sampling (cluster sampling, for instance, reduces precision). A design effect less than 1.0 means the sampling design increases precision compared to simple random sampling (stratified sampling, for instance, increases precision).