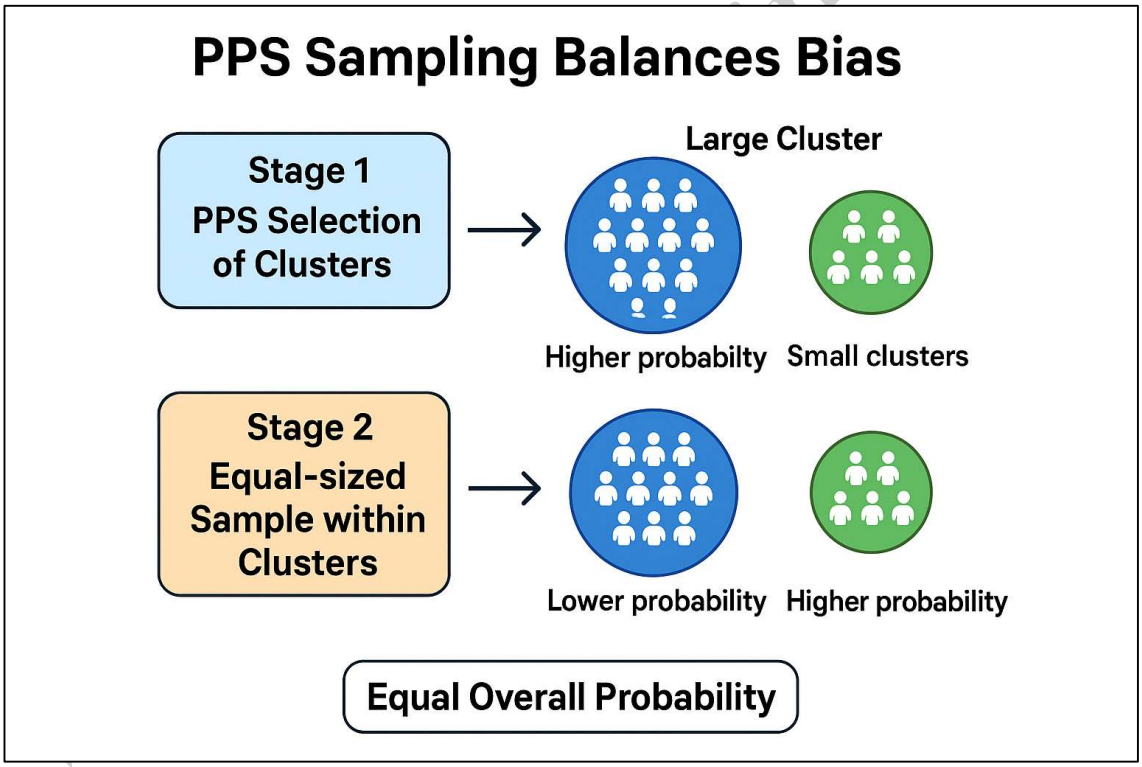
Probability Proportional to Size (PPS) Sampling – A Complete Guide

When we do surveys, not all groups in a population are the same size. Some districts have thousands of households, while others have just a few. If we treated every district equally, people in smaller ones could be overrepresented.

Probability Proportional to Size (PPS) sampling fixes this problem. Bigger clusters (like large districts or villages) get a higher chance of being picked, while smaller ones get a lower chance. But here's the trick: once the second stage of sampling is done, every household ends up with the same overall chance of being selected.

That's why PPS is so widely used in big national surveys — it's fair, efficient, and makes analysis much simpler.

The method was first introduced by **Hansen and Hurwitz (1943)** in the context of a **two-stage sampling framework with replacement**.



How PPS Sampling Works

Step 1 – Define the Target Population

Clearly specify the population of interest and identify the **primary sampling units (PSUs)**, such as districts, villages, or blocks.

Step 2 – Determine the Sampling Frame

Prepare a complete list of all PSUs, including their **measure of size (MOS)**—such as population, number of households, revenue, etc.

Step 3 – Assign Selection Probabilities

For each PSU:

$$\text{Prob1} = \frac{\text{Cluster Population}}{\text{Total Population}}$$

Interpretation:

- This is the probability that a **cluster (PSU)** (e.g., district, village, block) is selected in the **first stage unit**.
- Larger clusters (with more people) have higher **Prob1**, because they represent a larger share of the total population.
- Example: Assume total households in a state = 1,00,000.
- Large district: 20,000 households →

$$\text{Prob1} = \frac{20{,}000}{1{,}00{,}000} = 0.20$$

- Meaning this district has a 20% chance of being selected in the first stage.
- Small district: 5,000 households →

$$\text{Prob1} = \frac{5{,}000}{1{,}00{,}000} = 0.05$$

- Meaning this district has a 5% chance of being selected in the first stage.

Step 4 – Select PSUs (Stage 1)

PSUs can be selected in two main ways:

- **Cumulative Total Method** (using simple random sampling (SRS) or systematic sampling), or
- **Lahiri's Method** (select a unit by SRS and then accept/reject based on its size).

Step 5 – Select Units within Clusters (Stage 2)

From each selected PSU, select the **same number** of secondary sampling units (SSUs)—households, individuals, etc.—using **SRS**.

For each unit in the cluster:

$$\text{Prob2} = \frac{\text{Number of Units Selected in Cluster}}{\text{Cluster Population}}$$

Interpretation:

- This is the probability that an **individual (SSU)** within a selected cluster is chosen.
- Since the same number of individuals are selected from each cluster (say 100 households per district), units in **smaller clusters have a higher within-cluster chance of selection**, while units in larger clusters have a lower chance.
- Example: Larger district: 20,000 households; sample 100 households

$$\text{Prob2} = \frac{100}{20{,}000} = 0.005$$

- Meaning each household in that district has 0.5% chance to be selected in the second stage unit.
- Smaller district: 5,000 households; sample 100 households

$$\text{Prob2} = \frac{100}{5{,}000} = 0.02$$

- So, within-cluster selection probability is higher in the smaller district (2%) than in the larger one (0.5%).

Step 6 – Calculate Overall Probability and Weight

The overall probability of selecting an individual is the product of first-stage and second-stage probabilities:

$$\text{Overall Probability} = \text{Prob1} \times \text{Prob2}$$

Interpretation:

- This is the probability that a **specific individual in the entire population** is ultimately selected.
- Multiplying first-stage and second-stage probabilities balances the bias:
  - Larger clusters → higher chance in Stage 1 but lower within-cluster chance in Stage 2.
  - Smaller clusters → lower chance in Stage 1 but higher within-cluster chance in Stage 2.
- Result: Every individual, regardless of cluster size, has the same overall probability of selection.
- For larger district

$$\text{Overall Probability} = 0.20 \times 0.005 = 0.001$$

- For smaller district

$$\text{Overall Probability} = 0.05 \times 0.02 = 0.001$$

- Overall probability is equal (0.001) in both cases → self-weighting.

Finally, the **sampling weight** for each selected individual is simply the **reciprocal of the overall probability**.

$$\text{Weight} = \frac{1}{\text{Overall Probability}}$$

This weight reflects how many households in the population each sampled household represents. Since the overall probability of selection is the same for every household (self-weighting property of PPS), the weight is also the same for all households.

Example:

Suppose the Overall Probability of selecting any household = 0.001 (0.1%).

Then,

$$\text{Weight} = \frac{1}{0.001} = 1{,}000$$

This means each household in the sample stands for 1,000 households in the whole population.

Why is Weight Important?

- Weights are used to inflate the sample data back to the population level.
- For instance, if we survey 100 households with a weight of 1000 each, our data represents 100 × 1000 = 1,00,000 households in the population.
- It ensures that survey estimates (like averages, proportions, totals) are unbiased and representative of the population.

Why PPS Creates a Self-Weighting Sample

Stage 1 (PPS): Large clusters are more likely to be selected.

Stage 2 (Equal Sample per Cluster): Units within small clusters have higher within-cluster selection probabilities.

Balance: These two effects cancel each other, making the overall selection probability equal for all individuals, regardless of cluster size.

Thus, PPS yields a self-weighting design, which is why it is extensively used in national household surveys, demographic studies, and large-scale health or education research.

Methods for PPS Selection:

When we use Probability Proportional to Size (PPS) sampling, there are two common ways to select clusters (PSUs). Both aim to give larger clusters a higher chance of being picked, but they work differently.

Cumulative Total Method: This is the most widely used method for PPS.

Steps:

1. **List All Clusters with Their Sizes**: Prepare a list of all clusters along with their respective sizes (e.g., population, households etc.).
2. **Compute the Cumulative Total of Sizes**: Add cluster sizes sequentially
3. **Select Clusters Proportional to Size**: We can pick PSUs (clusters) using either **Simple Random Sampling (SRS)** or **Systematic Sampling**:
    - In Simple Random Sampling (SRS), we generate random numbers between 0 and the total cumulative size of the population, and each number falls into a range that corresponds to a cluster. For instance, imagine there are three villages with household sizes of 200, 300, and 500. Their cumulative totals are 200, 500, and 1,000. If we generate a random number such as 450, it falls within the cumulative range of 201–500, which corresponds to Village 2.

Therefore, Village 2 is selected. This process is repeated until the required number of primary sampling units is obtained.

- Systematic Sampling: Calculate the sampling interval

$$\text{Sampling Interval (SI)} = \frac{\text{Total Population}}{\text{Number of clusters needed}}$$

Pick a random start between 0 and SI. Keep adding the interval to the random start to get the selection points. Each selection point falls into a cumulative total range, which identifies the selected cluster.

For example, suppose the total population is 1,000 households and we need to select 2 villages. The sampling interval (SI) is calculated as 1,000 ÷ 2 = 500. Next, we pick a random start between 0 and 500; let's say the random start is 300. The selection points will then be 300 and 800 (by adding the interval of 500). Now, consider the cumulative household totals of the villages: Village 1 = 200, Village 2 = 500, and Village 3 = 1,000. The first selection point, 300, falls in the cumulative range of Village 2 (201–500), while the second selection point, 800, falls in the range of Village 3 (501–1,000). Therefore, the two selected villages are Village 2 and Village 3.

4. **Select a Fixed Number of Individuals from Each Selected Cluster**: Once clusters are selected, a **fixed number of individuals** is sampled from each cluster (using SRS within the cluster). This ensures that all individuals across the population have an **equal probability of selection**, even though clusters were chosen with probabilities proportional to their size.

Lahiri's Method: This is a simpler, field-friendly approach — especially useful when cumulative totals aren't pre-computed.

Steps:

1. **List All Clusters with Their Sizes**: Prepare a complete list of all clusters (PSUs) along with their respective sizes (such as population, number of households, etc.).
2. **Randomly Select a Cluster**: Select a cluster uniformly at random from the list, without considering size at this step.
3. **Generate a Random Number for Size Comparison:** For the selected cluster, generate a random number between 0 and the maximum cluster size in the population.
4. **Acceptance or Rejection:** Compare the random number with the actual size of the selected cluster:
    a. If the random number ≤ size of the selected cluster, accept the cluster.
    b. If the random number > size of the selected cluster, reject the cluster and return to Step 2.
5. **Repeat:** Continue this process until the desired number of clusters.

Suppose the largest cluster has 500 households. You randomly pick Village B, which has 300 households. Then you generate a random number. If the random number is 250, it falls within Village B's size (0–300), so you ✅ accept Village B. If the random number is 400, which is greater than 300, you ❌ reject Village B and repeat the process.

Village	Size (Households)	Acceptance Range
A	200	0–200
B	300	0–300
C	500	0–500

Larger villages are more likely to be accepted because their size covers a wider range of possible random numbers. This makes Lahiri's method naturally proportional to size, even without calculating cumulative totals.

Application of PPS Sampling using both methods:

To make this concept crystal clear, let's walk through an example using population data from the Indian state of **Andhra Pradesh**, which has 26 districts. Suppose we want to select **10 districts** (clusters) in the **first stage**, and then sample **1,000 individuals** (or households, depending on the study) from each selected district in the **second stage**.

Cumulative Total Method:

Step1: List Districts and Population: We start with population data for each district, calculate the **cumulative population**, and determine the **cumulative range** for each district. This range will help us determine which district a randomly selected number belongs to. Total population of Andhra Pradesh = **5,14,07,143.** We want to draw a sample of 10 districts using PPS sampling**.**

District-wise Cumulative Population Table

Sl. No.	District	Population	Cumulative Population	Cumulative Range	Prob 1
1	Srikakulam	21,91,471	21,91,471	1 - 21,91,471	0.04
2	Parvathipuram Manyam	9,25,340	31,16,811	21,91,472 - 31,16,811	0.02
3	Vizianagaram	19,30,811	50,47,622	31,16,812 - 50,47,622	0.04
4	Alluri Sitharama Raju	9,53,960	60,01,582	50,47,623 - 60,01,582	0.02
5	Visakhapatnam	19,59,544	79,61,126	60,01,583 - 79,61,126	0.04
6	Anakapalli	17,26,998	96,88,124	79,61,127 - 96,88,124	0.03
7	Kakinada	20,92,374	1,17,80,498	96,88,125 - 1,17,80,498	0.04
8	East Godavari	18,32,332	1,36,12,830	1,17,80,499 - 1,36,12,830	0.04
9	Konaseema	17,19,093	1,53,31,923	1,36,12,831 - 1,53,31,923	0.03
10	Eluru	20,06,737	1,73,38,660	1,53,31,924 - 1,73,38,660	0.04
11	West Godavari	17,79,935	1,91,18,595	1,73,38,661 - 1,91,18,595	0.03
12	NTR	22,18,591	2,13,37,186	1,91,18,596 - 2,13,37,186	0.04
13	Krishna	17,35,079	2,30,72,265	2,13,37,187 - 2,30,72,265	0.03
14	Guntur	20,91,075	2,51,63,340	2,30,72,266 - 2,51,63,340	0.04
15	Bapatla	15,80,000	2,67,43,340	2,51,63,341 - 2,67,43,340	0.03

Sl. No.	District	Population	Cumulative Population	Cumulative Range	Prob 1
16	Palnadu	20,41,723	2,87,85,063	2,67,43,341 - 2,87,85,063	0.04
17	Prakasam	17,35,000	3,05,20,063	2,87,85,064 - 3,05,20,063	0.03
18	Nellore	24,70,000	3,29,90,063	3,05,20,064 - 3,29,90,063	0.05
19	Kurnool	40,53,463	3,70,43,526	3,29,90,064 - 3,70,43,526	0.08
20	Nandyal	18,00,000	3,88,43,526	3,70,43,527 - 3,88,43,526	0.04
21	Anantapur	40,81,148	4,29,24,674	3,88,43,527 - 4,29,24,674	0.08
22	Sri Sathya Sai	18,00,000	4,47,24,674	4,29,24,675 - 4,47,24,674	0.04
23	YSR Kadapa	28,82,469	4,76,07,143	4,47,24,675 - 4,76,07,143	0.06
24	Annamayya	11,00,000	4,87,07,143	4,76,07,144 - 4,87,07,143	0.02
25	Chittoor	13,00,000	5,00,07,143	4,87,07,144 - 5,00,07,143	0.03
26	Tirupati	14,00,000	5,14,07,143	5,00,07,144 - 5,14,07,143	0.03
Total		5,14,07,143			1.00

Step 2: Select Clusters

SRS: We generate 10 random numbers between 1 and 5,14,07,143 (can be done using Excel or a random number table).

Sl. No.	Random Numbers	Cumulative Range	District selected	SRSWR	SRSWOR
1	38947079	3,88,43,527 - 4,29,24,674	District 21 (Anantapur)	✓	✓
2	43783077	4,29,24,675 - 4,47,24,674	District 22 (Sri Sathya Sai)	✓	✓
3	19308536	1,91,18,596 - 2,13,37,186	District 12 (NTR)	✓	✓
4	7516898	60,01,583 - 79,61,126	District 5 (Visakhapatnam)	✓	✓
5	47541127	4,47,24,675 - 4,76,07,143	District 23 (YSR Kadapa)	✓	✓
6	43379533	4,29,24,675 - 4,47,24,674	District 22 (Sri Sathya Sai)	✓	✗ (Already selected)
7	14154372	1,36,12,831 - 1,53,31,923	District 9 (Konaseema)	✓	✓
8	5612092	50,47,623 - 60,01,582	District 4 (Alluri Sitharama Raju)	✓	✓
9	23562935	2,30,72,266 - 2,51,63,340	District 14 (Guntur)	✓	✓
10	1437703	1 - 21,91,471	District 1 (Srikakulam)	✓	✓
11	33841961	3,29,90,064 - 3,70,43,526	District 19 (Kurnool)		✓ (Added to complete 10 unique districts)

✓ SRSWR allows repetition (District 22 repeated). ✓ SRSWOR avoids repetition and substitutes another district (District 19 added).

Systematic Sampling:

Step-by-Step Procedure

1. Determine Total Population: The total population across all 26 districts is: 5,14,07,143.

2. Calculate the Sampling Interval (k):

$$\text{Sampling Interval (k)} = \frac{\text{Total Population}}{\text{Sample Size}} = \frac{5,14,07,143}{10} = 51,40,714$$

So, we will select a district approximately every 51,40,714 units of population.

3. Choose a Random Start (R): We select a random number R between 1 and 51,40,714. Suppose R = 48,12,099.
4. Determine the Selection Points: We now calculate 10 selection points:

$$R, R + k, R + 2k, \cdots, R + (10 - 1)k$$

5. Match Each Selection Point to Districts: Each of these 10 selection points are then checked against the cumulative population ranges of the 26 districts to determine which districts are included in the final sample. It will be clear from the following table:

Sl. No.	Selection Point	Cumulative Range	Selected District
1	4,812,099	3,116,812 – 50,47,622	Vizianagaram (District 3)
2	9,952,813	79,61,127 – 96,88,124	Anakapalli (District 6)
3	15,093,527	1,36,12,831 – 1,53,31,923	Konaseema (District 9)
4	20,234,241	1,91,18,596 – 2,13,37,186	NTR (District 12)
5	25,374,955	2,67,43,341 – 2,87,85,063	Palnadu (District 16)
6	30,515,669	3,29,90,064 – 3,70,43,526	Kurnool (District 19)
7	35,656,383	4,29,24,675 – 4,47,24,674	Sri Sathya Sai (District 22)
8	40,797,097	4,76,07,144 – 4,87,07,143	Annamayya (District 24)
9	45,937,811	5,00,07,144 – 5,14,07,143	Tirupati (District 26)
10	51,078,525	5,00,07,144 – 5,14,07,143	Tirupati (District 26) *

Note: If Tirupati (District 26) is repeated, we have two options. With Replacement (SRSWR): We keep Tirupati as it is. Without Replacement (SRSWOR): We skip the repeated district and move to the next available district in sequence. Since there is no district after Tirupati (26th), we wrap around and select the first district, Srikakulam (District 1).

Lahiri's method:

1. Identify the Maximum Size (M): Determine the largest population among all 26 districts. For Andhra Pradesh:

$$M = 40,81,148\ (Anantapur)$$

2. Select a District at Random: Randomly select a district number between 1 and 26. Suppose we randomly pick District 5: Visakhapatnam (Population = 19,59,544)
3. Generate a Random Number (R): Generate a random number R between 1 and M (40,81,148). Suppose we get R=15,00, 000.

4. Check the Condition: Compare the random number with the population of the selected district: Since R = 15,00,000 ≤ 1 9,59,544, Visakhapatnam is selected.
5. If R had been greater than 19,59,544, the district would be rejected, and we would repeat the process with a new random district and a new random number.
6. Repeat Until Desired Sample Size is Reached: Continue this process until you select 10 unique districts (or however many are needed for your study).

Trial	Random District Number	District	Random Number (R)	Population	Selected?
1	5	Visakhapatnam	15,00,000	19,59,544	Yes
2	19	Kurnool	38,00,000	40,53,463	Yes
3	3	Vizianagaram	20,00,000	19,30,811	No (repeat)
4	3	Vizianagaram	12,00,000	19,30,811	Yes
5	21	Anantapur	39,00,000	40,81,148	Yes
6	7	Kakinada	25,00,000	20,92,374	No (repeat)
7	7	Kakinada	19,00,000	20,92,374	Yes
8	12	NTR	10,00,000	22,18,591	Yes
9	1	Srikakulam	5,00,000	21,91,471	Yes
10	23	YSR Kadapa	27,00,000	28,82,469	Yes
11	14	Guntur	20,00,000	20,91,075	Yes
12	24	Annamayya	11,00,000	11,00,000	Yes

Key Observation

- Some districts required multiple attempts (e.g., Vizianagaram and Kakinada) because the random number exceeded their population in the first attempt.
- Larger districts like Anantapur and Kurnool have a naturally higher chance of selection.

Second-Stage Sampling (Within Selected Districts):

Once districts (PSUs) are chosen, we move to the **second stage**. Here, we **do not sample in proportion to size**. Instead, we select a **fixed number of SSUs (say 1,000)** within each district using **Simple Random Sampling (SRS)**.

- These SSUs could be **households, individuals, schools, or enterprises**, depending on the study's focus.
- Since we select the **same number (1,000)** from each chosen district, the design becomes **self-weighting** at the SSU level: every individual (or household) in the state ultimately has an **equal chance of selection**.

Here's an illustration with **10 selected districts** using SRSWR and equal secondary samples:

Sl. No.	District selected	Cluster Population	Prob 1	SRSWR	Prob 2	Overall Probability	Overall Weight
1	District 21 (Anantapur)	40,81,148	8%	✓	0.02%	0.002%	50,000

Sl. No.	District selected	Cluster Population	Prob 1	SRSWR	Prob 2	Overall Probability	Overall Weight
2	District 22 (Sri Sathya Sai)	18,00,000	4%	✓	0.06%	0.002%	50,000
3	District 12 (NTR)	22,18,591	4%	✓	0.05%	0.002%	50,000
4	District 5 (Visakhapatnam)	19,59,544	4%	✓	0.05%	0.002%	50,000
5	District 23 (YSR Kadapa)	28,82,469	6%	✓	0.03%	0.002%	50,000
6	District 22 (Sri Sathya Sai)	18,00,000	4%	✓	0.06%	0.002%	50,000
7	District 9 (Konaseema)	17,19,093	3%	✓	0.06%	0.002%	50,000
8	District 4 (Alluri Sitharama Raju)	9,53,960	2%	✓	0.10%	0.002%	50,000
9	District 14 (Guntur)	20,91,075	4%	✓	0.05%	0.002%	50,000
10	District 1 (Srikakulam)	21,91,471	4%	✓	0.05%	0.002%	50,000

This is the essence of **self-weighting two-stage PPS sampling** – the backbone of large-scale surveys like **NFHS, NSSO, DHS**.

PPS Sampling with Rural–Urban Representation: Andhra Pradesh Example

In large-scale surveys, it is often necessary to ensure **both rural and urban populations are properly represented**. Andhra Pradesh has a mix of rural and urban households, and the sample should reflect this balance.

Step 1 – Rural–Urban Distribution in Andhra Pradesh

According to Census 2011 (latest full data):

- **Rural Population** = 3,47,77,389 (≈ 68%)
- **Urban Population** = 1,66,29,754 (≈ 32%)
- **Total Population** = 5,14,07,143

So, the rural–urban ratio in the state is **68:32**.

If we plan to sample **10,000 households statewide** (1,000 households each from 10 selected districts):

- Rural sample = 68% × 10,000 = 6,800 households
- Urban sample = 32% × 10,000 = 3,200 households

Step 2 – First Stage (District Selection using PPS)

Districts are selected with probability proportional to population size (as already demonstrated with PPS). Suppose the following **10 districts** are selected:

- Srikakulam, Visakhapatnam, Guntur, Anantapur, YSR Kadapa, Kurnool, Konaseema, Alluri Sitharama Raju, NTR, and Sri Sathya Sai.

Step 3 – Allocate Rural–Urban Sample within Each District

For each selected district, we look at its **rural vs. urban share** (as per Census data). Then, we split the 1,000 households allocated to each district accordingly.

Example calculations (illustrative, using approximate Census 2011 ratios):

District	Total Population	Rural %	Urban %	District Sample (1,000)	Rural Sample	Urban Sample
Srikakulam	21,91,471	83%	17%	1,000	830	170
Visakhapatnam	19,59,544	45%	55%	1,000	450	550
Guntur	20,91,075	67%	33%	1,000	670	330
Anantapur	40,81,148	72%	28%	1,000	720	280
YSR Kadapa	28,82,469	65%	35%	1,000	650	350
Kurnool	40,53,463	73%	27%	1,000	730	270
Konaseema	17,19,093	85%	15%	1,000	850	150
Alluri Sitharama Raju	9,53,960	90%	10%	1,000	900	100
NTR	22,18,591	60%	40%	1,000	600	400
Sri Sathya Sai	18,00,000	70%	30%	1,000	700	300
Total	—	—	—	**10,000**	**7,100**	**2,900**

This allocation closely respects the **statewide 68:32 rural–urban ratio**, while ensuring every selected district contributes proportionately.

Step 4 – Second Stage (Household Selection within Rural & Urban Strata)

- From each **district's rural frame** (list of villages/households), draw the rural sample using SRS.
- From each **district's urban frame** (list of wards/households), draw the urban sample using SRS.

Step 5 – Probability and Weighting

- **Stage 1 (Districts)**: Selected with PPS.
- **Stage 2 (Households)**: Selected in proportion to rural–urban share.

Thus, the **overall probability of selection remains equal** for every household statewide → **self-weighting sample**.

If, instead, you oversample urban areas (e.g., 50% rural and 50% urban sample for analytical purposes), you must apply **weights** in analysis to restore the true 68:32 population ratio.

Conclusion:

This PPS with rural–urban allocation ensures **both representativeness and efficiency**. It preserves the **self-weighting design**, unless deliberate oversampling is introduced (in which case, weights correct for the imbalance).