

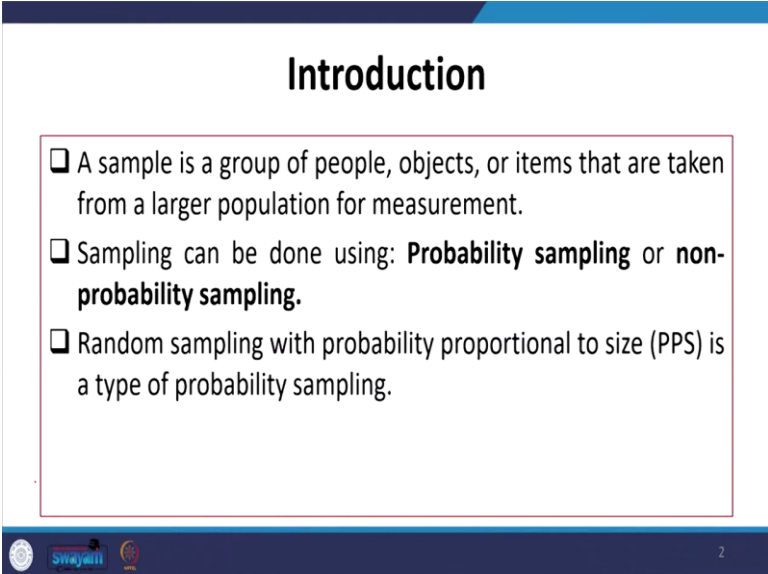
Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 09
Sample Size Determination and Probability Proportional to Size Sampling

Welcome friends to the NPTEL MOOC module on handling the healthcare survey dataset. We are on the verge of 2nd week and we are trying to explain the requirement of the field survey for health care research. We are trying to understand the type of sampling and sample size determination. The sample size determination as we already pointed out in the last lecture is very important.

And then it connects to your margin error, it connects to your standard deviation or it connects to your confidence level. So, those clarifications, I have already given in the previous lecture. Now we are concentrating on sample size determination and a special technique nowadays utilized is called probability proportional to size sampling.

(Refer Slide Time: 01:35)



Introduction

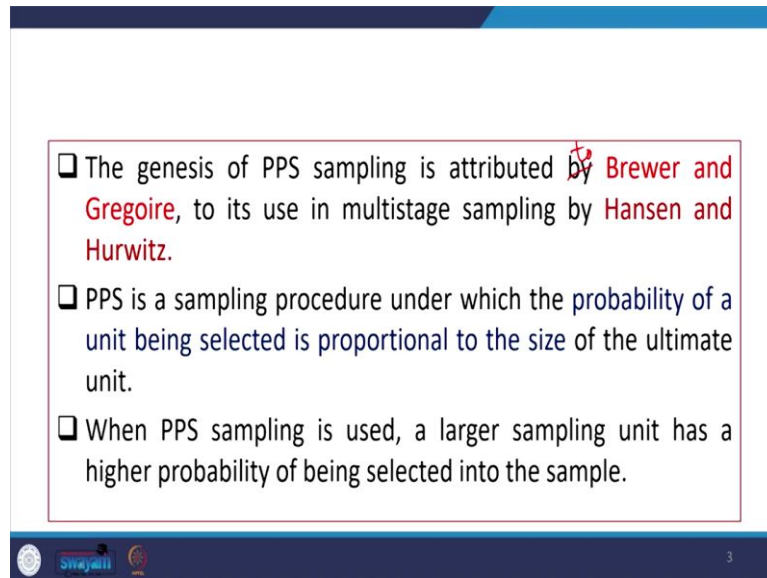
- A sample is a group of people, objects, or items that are taken from a larger population for measurement.
- Sampling can be done using: **Probability sampling** or **non-probability sampling**.
- Random sampling with probability proportional to size (PPS) is a type of probability sampling.

swayamii 2

Now, in the very backdrop, I just wanted to mention that a sample is a group of people and objects or items that are taken from a large population for measurement. Sampling can be done using either the probability method or by non-probability sampling method. Random sampling is often used with probability proportional to size sampling.

And in fact, random sampling with probability proportional to size is a type of probability sampling. The genesis of PPS sampling is attributed to Brewer and Gregoire.

(Refer Slide Time: 02:19)



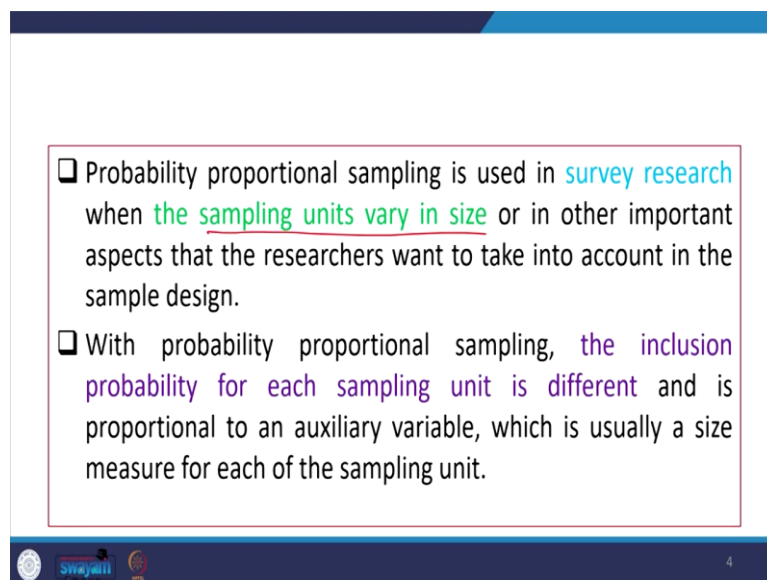
Slide 3 contains three bullet points:

- ❑ The genesis of PPS sampling is attributed by Brewer and Gregoire, to its use in multistage sampling by Hansen and Hurwitz.
- ❑ PPS is a sampling procedure under which the probability of a unit being selected is proportional to the size of the ultimate unit.
- ❑ When PPS sampling is used, a larger sampling unit has a higher probability of being selected into the sample.

The slide footer includes the Swayam logo and the number 3.

So, its use in multistage sampling by Hansen and Hurwitz. PPS is a sampling procedure under which the probability of a unit being selected and is proportional to the size of the ultimate unit. When PPS sampling is used a larger sampling unit has a higher probability of being selected for the sampling.

(Refer Slide Time: 02:44)



Slide 4 contains two bullet points:

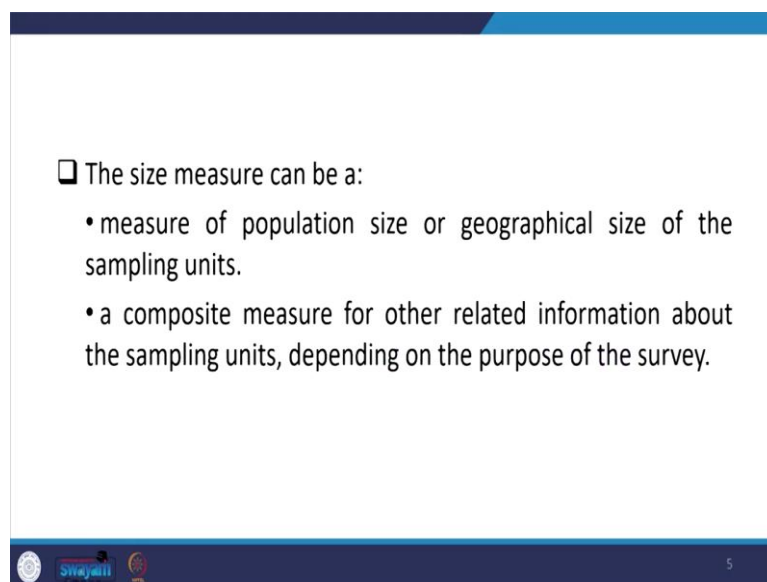
- ❑ Probability proportional sampling is used in survey research when the sampling units vary in size or in other important aspects that the researchers want to take into account in the sample design.
- ❑ With probability proportional sampling, the inclusion probability for each sampling unit is different and is proportional to an auxiliary variable, which is usually a size measure for each of the sampling unit.

The slide footer includes the Swayam logo and the number 4.

Probability proportional sampling is used in survey research when the sampling units vary in size. So, this is important when the sampling units are very heterogeneous, the probability sampling technique is suggested. Other important aspects that the researchers want to take into account in the sample design as well in the PPS technique.

With the probability proportional sampling, the inclusion probability for each sampling unit is different and is proportional to an auxiliary variable, which is usually a size measure for each sampling unit.

(Refer Slide Time: 03:30)



□ The size measure can be a:

- measure of population size or geographical size of the sampling units.
- a composite measure for other related information about the sampling units, depending on the purpose of the survey.

At the bottom of the slide, there is a blue footer bar containing a logo on the left, the text 'Sivajali' in the center, and the number '5' on the right.

The size measure can be a measure of population size or geographical area of the sampling unit. A composite measure of other related information about the sampling units depending on the purpose of the survey is also considered in the PPS.

(Refer Slide Time: 03:51)

- ❑ PPS is widely used in survey research with multistage design.
- ❑ Many large-scale social surveys with complex sample design have implemented probability proportional sampling in at least one of the sampling stages.
- ❑ Example: National Family Health Survey, National Sample Survey Office, Health and Retirement Survey

PPS is widely used in survey research with multi-stage design. Many large-scale social surveys with complex sample designs have implemented probability proportional to size sampling in at least one of the sampling stages. For example, PPS is used in the national family health survey (NFHS), National Sample Survey (NSS) and also in health retirement survey.

(Refer Slide Time: 04:16)

DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH BASED ON PRECISION RATE AND CONFIDENCE LEVEL

❑ **Sample size when estimating a mean:** The confidence interval for the universe mean, μ , is given by

$$\mu = \bar{X} \pm z \frac{\sigma_p}{\sqrt{n}}$$

acceptable error $e = z \cdot \frac{\sigma_p}{\sqrt{n}}$

$$n = \frac{z^2 \sigma_p^2}{e^2}$$

where

- N = size of population
- n = size of sample
- e = acceptable error (the precision)
- σ_p = standard deviation of population
- z = standard variate at a given confidence level.

Now we are discussing about a bit on sample size what it requires and how it is relevant. What parameters are we urgently required to estimate? Like it is based on the extent of

precision in the result and the confidence level. The first aspect we are calculating is when the sample size is estimated given the fact that we are estimating at a mean level.

For sample size when estimating a mean, we require the confidence interval of the universe that is μ and μ at the mean level is defined as:

$$\mu = \bar{X} \pm z \frac{\sigma_p}{\sqrt{n}}$$

So, capital N is the size of the entire population and small n is the sample population and e is the acceptable error and which is also called the precision and σ_p is in fact, the standard deviation of the population and z is the tabulated value or the standard variate at a given confidence level. So, z is determined accordingly, now based on that and this is our population mean.

Based on this we can estimate our margin error that is:

$$e = z \cdot \frac{\sigma_p}{\sqrt{n}}$$

So, from there we can calculate the sample size. So, sample size is here this is basically:

$$n = \frac{z^2 \sigma^2}{e^2}$$

(Refer Slide Time: 07:12)

In case of finite population, the confidence interval for the universe mean, μ , is given by

$$\bar{X} \pm z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$$

acceptable error $e = z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$

$$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)e^2 + z^2 \sigma_p^2}$$

The slide contains the following text and formulas:

- In case of finite population, the confidence interval for the universe mean, μ , is given by
- $$\bar{X} \pm z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$$
- acceptable error $e = z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$
- $$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)e^2 + z^2 \sigma_p^2}$$

Arrows in the final formula indicate that $z \frac{\sigma_p}{\sqrt{n}}$ from the acceptable error formula is substituted into the sample size formula.

But when we have a finite population in that case we are supposed to multiply with this portion. This is

$$\sqrt{\frac{(N-n)}{(N-1)}}$$

So, accordingly, we have this component and from there we solve it:

$$n = \frac{z^2 \cdot N \cdot \sigma_p^{2*}}{(N-1)e^2 + z^2 \sigma_p^2}$$

(Refer Slide Time: 07:49)

Illustration

A researcher want to collect cross-sectional data on past alcohol habits and current diagnose of liver disease, Determine the size of the sample for universe with $N = 5000$ on the basis of the following information:

- 1) the variance = 4 g/dl on the basis of past records.
- 2) estimate should be within 0.8 g/dl of the true average weight with 99% probability.

Handwritten red annotations: 'z' next to '99%' and 'e' with an arrow pointing to '0.8 g/dl'.

Now one example we can cite here, one researcher wants to collect cross-sectional data on past alcohol habits and current diagnose of liver disease, determine the size of the sample for this universe of $N = 5000$ on the basis of the following information. So, we need to find out the sample size given the information that the variance is 4.

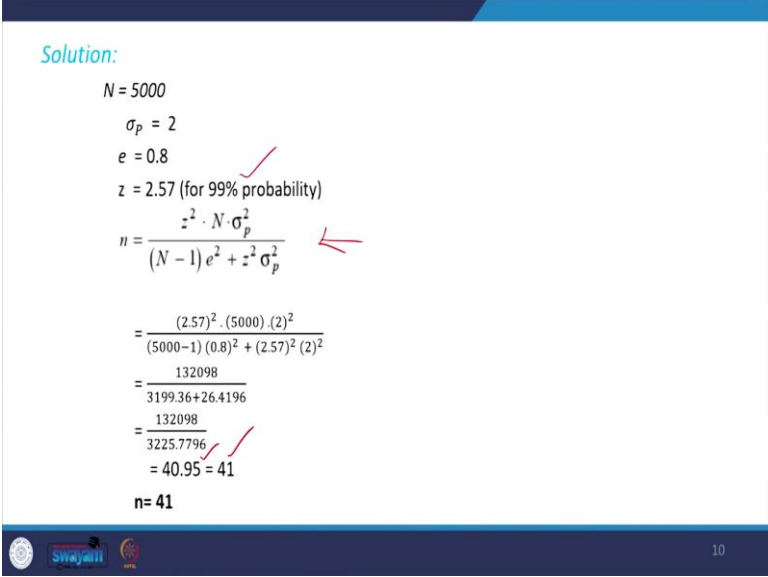
So, the standard deviation is the square root of 4 which is 2, on the basis of past records there is a variance of the distribution is given as 4. Our estimate should be within 0.8 margin error of the true weight with 99 percent probability. So Z value is 2.57 with a confidence level of 99 percent. The margin error (e) is 0.8 and our capital N is given. So, when your capital N is given; that means, you have to apply the second formula.

(Refer Slide Time: 09:12)

Solution:

$$N = 5000$$
$$\sigma_p = 2$$
$$e = 0.8$$
$$z = 2.57 \text{ (for 99\% probability)}$$
$$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)e^2 + z^2 \sigma_p^2}$$
$$= \frac{(2.57)^2 \cdot (5000) \cdot (2)^2}{(5000-1)(0.8)^2 + (2.57)^2 (2)^2}$$
$$= \frac{132098}{3199.36 + 26.4196}$$
$$= \frac{132098}{3225.7796}$$
$$= 40.95 = 41$$

n = 41



In that case, we are applying this formula which we have derived in the previous equation. So, the z value at 99 percent confidence level is 2.57. At 95 percent, the z value is 1.96 and the margin error is 0.8; that means, your result will be differentiated by point plus 0.8 and minus 0.8.

Your z value could be differentiated by this margin error. Once we put all the values in the formula we got a sample size n is equal to 40.95. So, in a round figure, it is 41.

(Refer Slide Time: 10:14)

Sample size when estimating percentage or proportion:

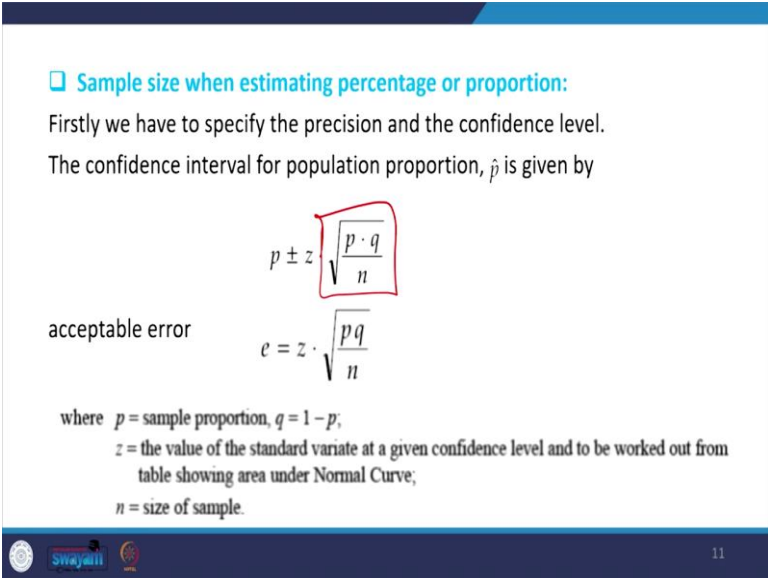
Firstly we have to specify the precision and the confidence level.
The confidence interval for population proportion, \hat{p} is given by

$$p \pm z \sqrt{\frac{p \cdot q}{n}}$$

acceptable error

$$e = z \cdot \sqrt{\frac{pq}{n}}$$

where p = sample proportion, $q = 1 - p$;
 z = the value of the standard variate at a given confidence level and to be worked out from table showing area under Normal Curve;
 n = size of sample.



If you are given information about proportion or the percentage then we have to specify the precision and confidence level. The confidence interval for a population proportion is given as:

$$\hat{P} = p \pm z \sqrt{\frac{pq}{n}}$$

From this, we can able to calculate the n value and this is also suggesting the fact that p is your probability of inclusion and q is non-inclusion.

(Refer Slide Time: 11:17)

Bryman 2016 Formula

$$n = \frac{z^2 \cdot p \cdot q}{e^2}$$

In case of finite population

$$n = \frac{z^2 \cdot p \cdot q \cdot N}{e^2 (N - 1) + z^2 \cdot p \cdot q}$$

12

Now another formula we have is called Bryman 2016 formula and this is based on proportion instead of the exact figures. Then look at in the place of sigma square we have actually taken p into q.

(Refer Slide Time: 12:25)

COCHRAN'S FORMULA FOR CALCULATING SAMPLE SIZE WHEN THE POPULATION IS INFINITE

Sample size to estimate a proportion

- ❑ A researcher want to check nature of treatment during covid-19 in a district. He asks, "How large a sample size do I need?"
- ❑ Required steps:
 - ❑ Margin Error: 2.5% in a 2-tailed test, generally
 - ❑ confidence intervals: 95% or different
 - ❑ Guesstimate of the proportion: $p=30\%$ who supports traditional medicine
- ❑ The margin error is 1.96 times the SE (Cochran (1977))

$$ME = z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

swayamii 13

We can calculate the sample size by Cochran's formula. When the population is infinite it has given another formula you can follow this example I am not emphasizing much on it.

If any proportion is not given some guess estimates could be taken like the p percentage. If you are testing how much sample size is taken to understand the nature of treatment during Covid-19. How much percentage researcher takes to check the nature of treatment during Covid-19 in a district?

So, it might be the case that the nature of treatment is taken by 30 percent of the population. Some guess estimates should be taken or some based on certain literature.

(Refer Slide Time: 13:44)

□ Z-score is 1.645 for 90%, 1.96 for 95%, 2.58 for 99%

$$n_0 = \frac{z^2 pq}{e^2}$$
$$0.025 = 1.96 \sqrt{\frac{0.3 \times 0.7}{n}}$$
$$\frac{0.3 \times 0.7}{n} = \left(\frac{0.025}{1.96}\right)^2 = .0001627$$
$$n = \frac{0.3 \times 0.7}{.0001627} = 1291$$

So we would need a sample of about 1300 people at 90% confidence interval, for which $z = 1.645$

$$0.025 = 1.645 \sqrt{\frac{0.3 \times 0.7}{n}}$$

we can quickly find $n = 909$

14

So, based on that the margin error is calculated likewise we did before. Now here is our z score is different if it is 95 or 90 percent, the z value is 1.645 or 1.96. So, first of all, we can find out the sample size based on the formula which I have suggested.

So, 1291 is the sample size. We would need around 1300 people at 90 percent level of confidence.

(Refer Slide Time: 14:59)

□ Having no idea of proportions

□ Assume 50%, i.e. $p^* = 0.5$

$$0.025 = 1.96 \sqrt{\frac{0.5 \times 0.5}{n}}$$
$$n = 1537$$

15

When we do not have any idea then 50 percent could be assumed. When your proportion of inclusion is higher you are supposed to take more portion to be surveyed.

(Refer Slide Time: 15:25)

COCHRAN'S FORMULA FOR CALCULATING SAMPLE SIZE WHEN THE POPULATION IS FINITE

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$


n_0 is the sample size derived from equation (Cochran 1977) and N is the population size

A researcher wants to assess the mean health expenditure for all homeowners in Rajasthan. A survey ten years ago got a sample mean and standard deviation of Rs.1400 and Rs.1000.

➤ How many household should be sampled for a 95% confidence interval to have a margin of error of Rs.100?

$$ME = t \frac{s}{\sqrt{n}}$$
$$n = \left(\frac{st}{ME} \right)^2$$

With $s = 1000, t = 1.96, ME = 100$, we get $n = 384$



Cochran's formula is used also when we have a finite population. In the previous example, we said about infinite population. When it is finite first of all we have to find out the value.

For example, a researcher wants to assess the mean health expenditure for all homeowners in Rajasthan. A survey 10 years ago got a sample mean and standard deviation of 1400 and 1000. How many households would be sampled with 95 percent confidence interval to have a margin error of 100?

Based on this formula we calculated n . In this case, our S is 1000 and the t value we have taken at 1.96 and margin error is 100 and we get n equal to 384. So, n is here the sample size derived from the equation of Cochran and N is the population size.

(Refer Slide Time: 17:34)

YAMANE (1967)

Sample sizes calculated by Yamane's formula

$$n = \frac{N}{1 + Ne^2}$$
$$n_i = \frac{N_i}{N} * n$$

n= the sample size
N= the population size and
e = the acceptable sampling error (5%)

Sl. no. of schools	Population size, N	Sample size, n for 95% confidence level:		
		±5%	±7%	±10%
1	450	212	136	82
2	582	229	150	85
3	693	254	158	87
4	799	266	163	89
5	806	267	163	89
6	845	272	164	89
7	858	273	165	90
8	892	276	166	90
9	909	278	167	90
10	922	279	167	90
11	9 85	285	169	91
12	1009	287	170	91

So, we are now going to calculate n which is equal to:

$$n = \frac{N}{1 + Ne^2}$$

as per the Yamane formula. What is the Ni? Ni, in fact, the sample size of this particular category is the particular ratio of samples from the group is taken. Now N is the total population size and e is the acceptable sample region.

As per the Yamane formula like you have population size if it is given in different numbers with 5 percent plus/minus (+/-) confidence interval. How they calculated the sample size is given. So, there is a different population, you can mention the population and you can able to calculate your individual N.

(Refer Slide Time: 18:50)

$$n = \frac{N}{1 + Ne^2}$$

n= the sample size
N= the population size (1072756 and 83260 households in rural and urban areas of two districts respectively) and
e = the acceptable sampling error (5% = 0.05)

Two districts of Odisha (i.e. Balasore and Mayurbhanj) will be selected for personal interview. ~~The rationale of the same is presented in the appendices.~~
The total number of urban households in Balasore district is 47360 and 35900 in Mayurbhanj district. The total number of sample selected from urban areas (Yamane formula) are $398.0875 = 400$ households and similarly another (397.87) 400 are selected from rural areas since the total number of rural households in these two districts is 1072756. Therefore, the total number of household selected for sample survey is 800.

As I will point out here in this example suppose we have the sample size (n), and the population size(N) is 1072756 and 8310286 households in rural-urban areas of two districts respectively and e is the acceptable sampling error of 5 percent that is 0.05.

In my own survey, two districts of Odisha were selected for a personal interview that is present in the appendices. But now I am not going to see the appendices. I can just simply mention it here the total number of urban households in Balasore is 47360.

The total number of samples selected from the urban areas as per the Yamane formula is 2400 households. Similarly, another 400 are selected from rural areas with this formula based on the two districts and their population size.

So, the total sample to be taken is 400+400 which is 800. So, the Yamane formula actually gives you a pretty easy way to find out the sample size. What are the thumb rules? In this case, the thumb rules are the larger is the sample size the more accurate your estimate and the sampling error is inversely related to the size of the sample. I have already said the size of standard error depends on the size of sample size and it varies inversely with the size of the sample.

(Refer Slide Time: 21:01)

THUMB RULES

- ❑ The larger the sample size, the more **accurate** your estimates (*estimate of the true population mean*)
- ❑ **Sampling error** is inversely related to the size of the sample
- ❑ The size of **S.E.**, depends upon the sample size to a great extent and it varies inversely with the size of the sample.
 - ❑ *If double reliability is required i.e., reducing S.E. to 1/2 of its existing magnitude, the sample size should be increased four-fold.*
- ❑ **The central limit theorem** assures that the sampling distribution of the mean approaches normal distribution as the sample size increases.
 - ❑ *This fact holds especially true for sample sizes over 30.*

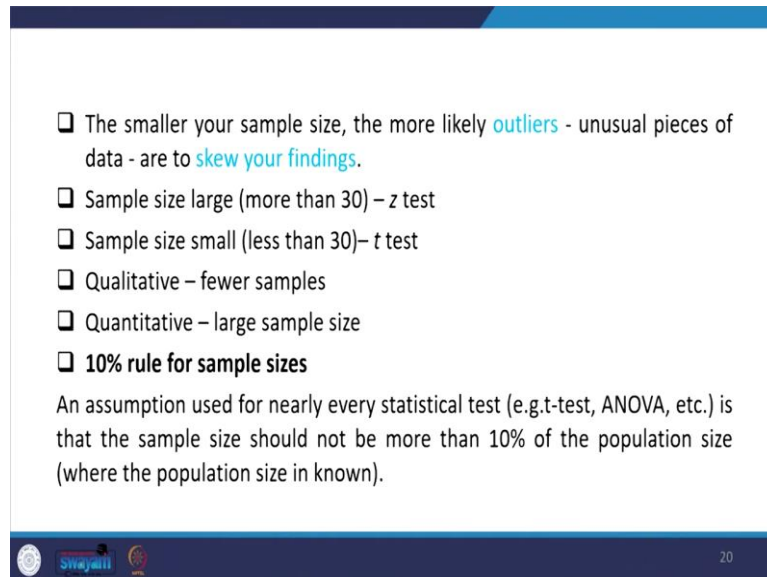
Swajati 19

If double reliability is required, then reducing samples error to half of it is an existing magnitude, the sample size should be increased fourfold.

What do you mean by a central limit theorem? This assures that the sampling distribution of the mean approaches the normal distribution as the sample size increases. This is what we will be drawing every time in our sample. It should be representative as suggested by the central limit theorem.

The fact holds especially true for the sample size is over 30. The smaller your sample size the more likely outliers you have on usual pieces of data and these are usually skewed sample size.

(Refer Slide Time: 22:06)



- ❑ The smaller your sample size, the more likely **outliers** - unusual pieces of data - are to **skew your findings**.
- ❑ Sample size large (more than 30) – z test
- ❑ Sample size small (less than 30)– t test
- ❑ Qualitative – fewer samples
- ❑ Quantitative – large sample size
- ❑ **10% rule for sample sizes**

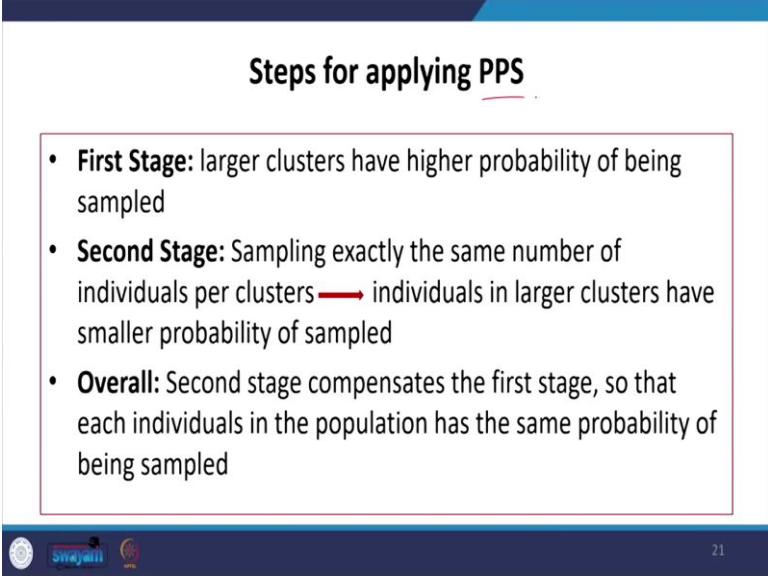
An assumption used for nearly every statistical test (e.g.t-test, ANOVA, etc.) is that the sample size should not be more than 10% of the population size (where the population size is known).

Swajati 20

If it is your studies qualitative fewer sample should be taken and if it is quantitative then larger samples will be taken. There should be a 10 percent thumb rule for sample sizes may be taken. But still, it is too high in number, an assumption used for nearly every statistical test (t-test, ANOVA) is that the sample size should not be more than 10 percent of the population size where the population size is known.

After saying all those about sample size determination. Now, we are categorically emphasizing on Probability Proportional to Size (PPS) sampling, there are some steps involved in the PPS.

(Refer Slide Time: 23:01)



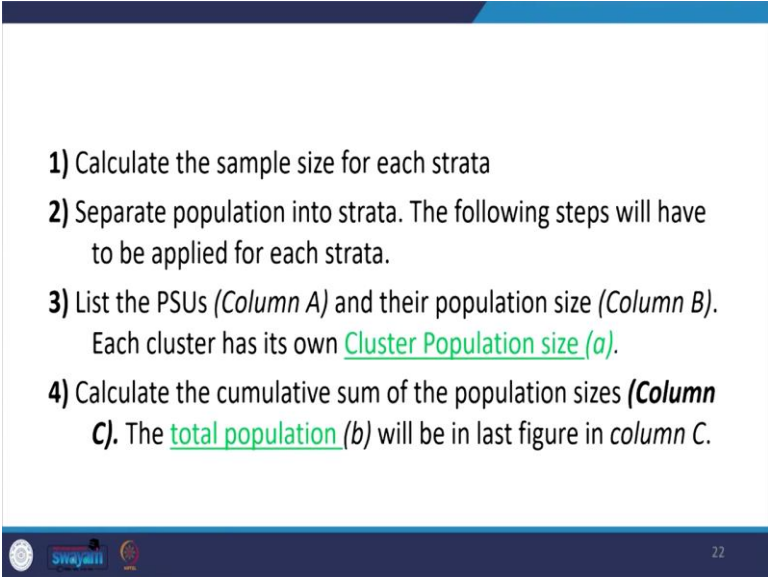
Steps for applying PPS

- **First Stage:** larger clusters have higher probability of being sampled
- **Second Stage:** Sampling exactly the same number of individuals per clusters → individuals in larger clusters have smaller probability of sampled
- **Overall:** Second stage compensates the first stage, so that each individuals in the population has the same probability of being sampled

21

The first stage is if you have larger clusters then higher the probability of being sampled. The second stage is sampling exactly the same number of individuals per cluster the individuals in larger clusters have a smaller probability of being sampled.

(Refer Slide Time: 23:27)



- 1) Calculate the sample size for each strata
- 2) Separate population into strata. The following steps will have to be applied for each strata.
- 3) List the PSUs (*Column A*) and their population size (*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 in last figure in *column C*.

22


Overall, the second stage compensates for the first stage. So, each individual in the population has the same probability of being sampled. We need to first calculate the sample size for each strata. Separate population into strata. The following steps will have to be applied for each strata. So, we need to list the Primary Sampling Units (PSUs).

(Refer Slide Time: 23:50)

5) Determine the Number of Cluster (d) that will be sampled in each strata.

6) Determine the Number of Individuals to be sampled from each cluster (c).

(Note: In order to ensure that all individuals in the population have the same probability of selection irrespective of size of their cluster, the same no. of individuals to be sampled from each cluster)




23

(Refer Slide Time: 23:51)

7) Divide the total population by the number of clusters to be sampled, to get the Sampling Interval (SI).

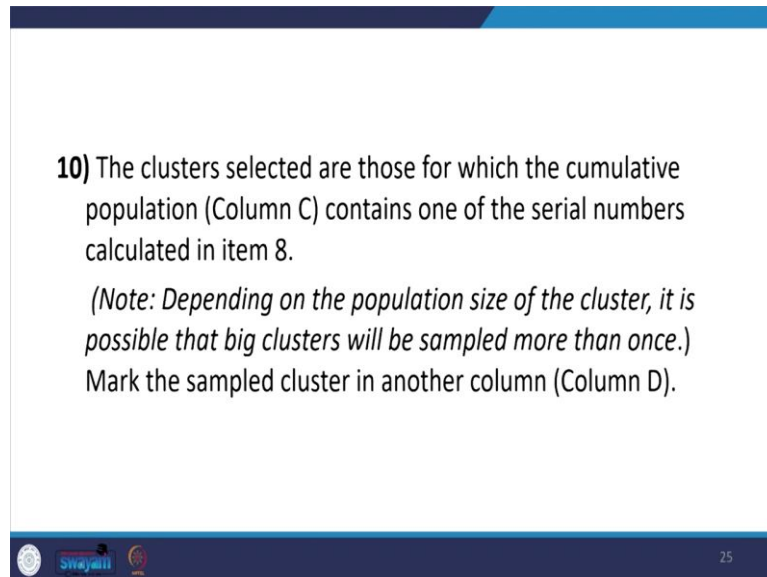
8) Choose a random no. between 1 and SI . This is the Random start (RS). The first cluster to be sampled contains this cumulative population (Column C).

9) Calculate the following series:
 $RS; RS+SI; RS+2SI; \dots RS+(d-1)*SI$



24

(Refer Slide Time: 23:51)



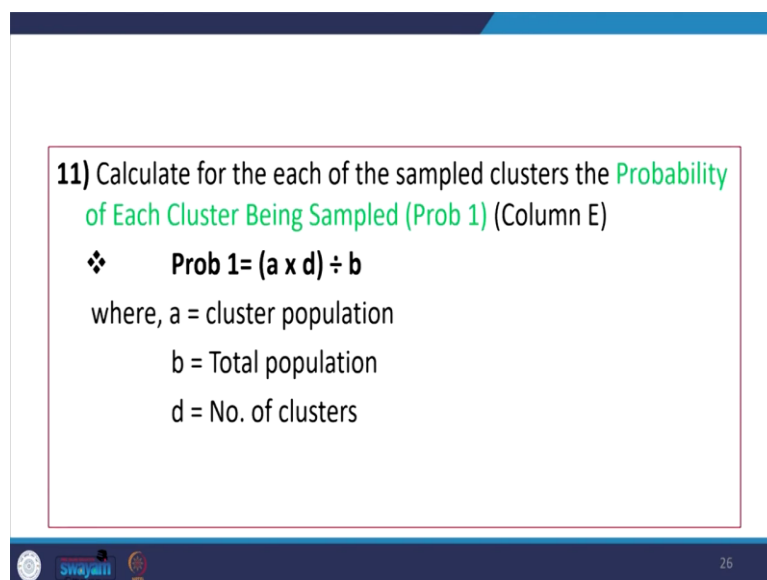
10) The clusters selected are those for which the cumulative population (Column C) contains one of the serial numbers calculated in item 8.

(Note: Depending on the population size of the cluster, it is possible that big clusters will be sampled more than once.)

Mark the sampled cluster in another column (Column D).

25

(Refer Slide Time: 23:52)



11) Calculate for the each of the sampled clusters the **Probability of Each Cluster Being Sampled (Prob 1)** (Column E)

❖ **Prob 1 = $(a \times d) \div b$**

where, a = cluster population
b = Total population
d = No. of clusters

26

You have to list the PSUs in column A and their population size in column B. I have one example for it I will show it to you then each cluster has its own population size each cluster wise we can calculate. Then you calculate the cumulative sum of the population sizes first we need to make them in order ascending or descending.

Then we make the cumulative population of that size. The next step is to calculate the cumulative sum of the population sizes that is the total population should be in column C, the cumulative population could be listed. We will determine the number of clusters;

how many clusters we need to make that will be sampled in each data. Then determine the number of individuals to be sampled from each cluster once you define the cluster.

Then individual sample could be taken from each cluster. The next step is to divide by the number of clusters to be sampled to get the sample intervals. How many sample you are supposed to collect choose a random number between 1 and the sample interval.

This is the Random start. Likewise, we follow systematic random sampling so the first clusters to be sampled contain the cumulative population in column C. Then you calculate the following series: RS ; $RS+SI$; $RS+2SI$;... $RS+(d-1) *SI$.

Finally, the clusters selected are those for which the cumulative population contains one of the serial numbers calculated in item number 8. Depending on the population size of the cluster it is possible that a big cluster will be sampled more than once and marked sample cluster in another column that we will mention.

Then we calculate for each of the sampled clusters the probability of each individual being sampled in each cluster. So, the probability is equal to c/a , a is a cluster population and c is the number of individuals to be sampled in each cluster. So, the probability of their inclusion could be also included.

(Refer Slide Time: 26:48)

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 clusters

27

(Refer Slide Time: 27:09)

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.

Basic Weight (BW)= 1/(prob 1*prob 2)

28

The number of individuals to be sampled in each cluster you can follow through with my example and I will be guiding it. So, we calculate the overall basic weight of an individual being sampled in the population. The weight is simply inverse to the probability of selection. $BW = 1/(prob1*prob2)$

(Refer Slide Time: 27:30)

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
- 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)					

29

Here is an example of how we followed. First of all, we make a group of the population out of the total. In India, clusters are blocks in districts. So, for each block let it be 1

cluster. In column c we have taken the cumulative sum, then you follow the c. The population is 20000 and clusters are 30.

Now, first of all, calculate probability 1 and probability 2. The sample is 3000 from 10 clusters using PPS. Then 4905 numbers could be taken based on the sample interval. Now I will guide it from another simplified data, but here you can just see how to calculate the probability of selection of each sample cluster.

We have mentioned a probability that is 51 percent for the 901 clustered sampled. So, basically probability 1 is equal to the probability of selection of each sampled cluster. Probability 2 is equal to the probability of selection of each individual in each of the sampled clusters. So, each individual is 300 if that number is allocated 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

So, now basically overall weight is the inverse of Prob1 into Prob2. So, Prob1 = 29% and Prob2 = 51%, this boils down to an overall weight of about 6.7 percent.

(Refer Slide Time: 30:52)

• No. of clusters (d) = **10**

• Sampling Interval = Cumulative population (B)/ Number of cluster (D)

= 20000/10

= **2000**

• Random Start (RS) = **905**

Now the number of clusters is 10 and the sample interval is equal to the cumulative population/Number of Cluster. So, 20000 divided by 10 that is 2000. So, the random start here is 905 out of 2000 population till 2000. We have got one random start that is 905, then 905 + 2000 is our 2905. This is plus 2000 is our 4905 similarly plus 2000 so

on. So, this is what all the numbers are selected. And now per individual cluster how many are selected that is as per the choice.

(Refer Slide Time: 32:02)

Slide 30 contains a list of calculations for cluster sampling. It starts with the number of clusters (d) = 10. Then it calculates the Sampling Interval as Cumulative population (B) divided by Number of cluster (D), which equals 20000/10 = 2000. Finally, it states the Random Start (RS) = 905. The slide has a blue header and footer with logos and the number 30.

- No. of clusters (d) = **10**
- Sampling Interval = Cumulative population (B)/ Number of cluster (D)
$$= 20000/10$$
$$= \mathbf{2000}$$
- Random Start (RS) = **905**

(Refer Slide Time: 32:03)

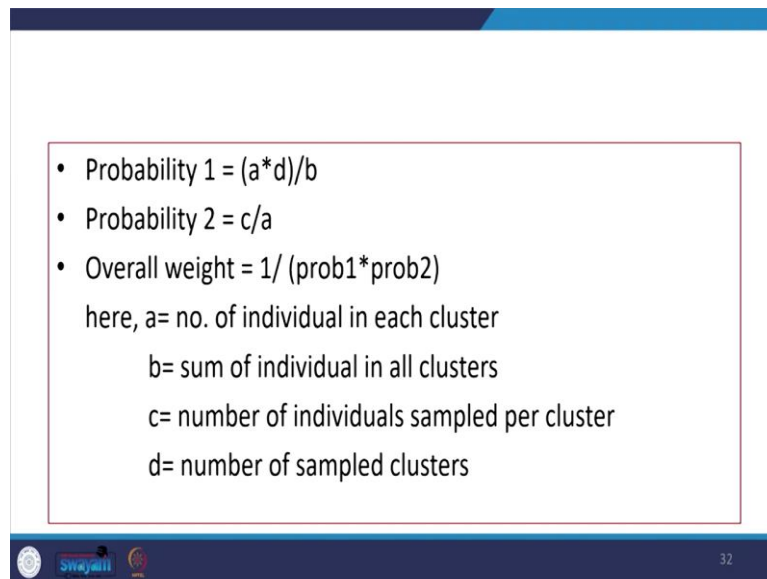
Slide 31 lists the series numbers for 10 clusters. Each number is calculated as RS + (i * SI), where i is the cluster number from 1 to 10. The numbers are: 1RS=905, 2RS+(1*SI)=2905, 3RS+(2*SI)=4905, 4RS+(3*SI)=6905, 5RS+(4*SI)=8905, 6RS+(5*SI)=10905, 7RS+(6*SI)=12905, 8RS+(7*SI)=14905, 9RS+(8*SI)=16905, and 10RS+(9*SI)=18905. Each number has a red checkmark next to it. The slide has a blue header and footer with logos and the number 31.

- Series numbers

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

Now I will tell you how it is selected. So, serial series numbers are like 905, and 2905. I think that is pretty clear to everyone.

(Refer Slide Time: 32:15)

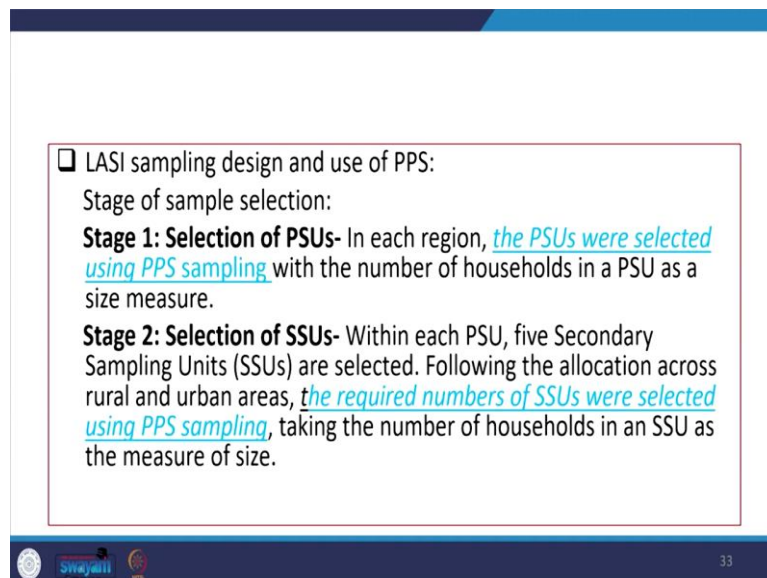


• Probability 1 = $(a*d)/b$
• Probability 2 = c/a
• Overall weight = $1/(\text{prob1}*\text{prob2})$
here, a= no. of individual in each cluster
b= sum of individual in all clusters
c= number of individuals sampled per cluster
d= number of sampled clusters

32

Probability 1 = $(a*d)/b$, Probability 2 = c/a and Overall weight = $1/(\text{prob1}*\text{prob2})$.

(Refer Slide Time: 32:27)



□ LASI sampling design and use of PPS:
Stage of sample selection:
Stage 1: Selection of PSUs- In each region, *the PSUs were selected using PPS sampling* with the number of households in a PSU as a size measure.
Stage 2: Selection of SSUs- Within each PSU, five Secondary Sampling Units (SSUs) are selected. Following the allocation across rural and urban areas, *the required numbers of SSUs were selected using PPS sampling*, taking the number of households in an SSU as the measure of size.

33

So, now we will discuss the LASI sampling design and use of PPS. You can follow on LASI database and LASI questionnaire which I have already discussed in the previous lectures. Then in stage one selection of PSUs, in each region, the PSUs were selected using PPS sampling with the number of households in a PSU as size of measure. Within each PSUs, five secondary sampling units (SSUs) are selected. Following the allocation

across rural and urban areas, the required number of SSUs were selected using PPS sampling.

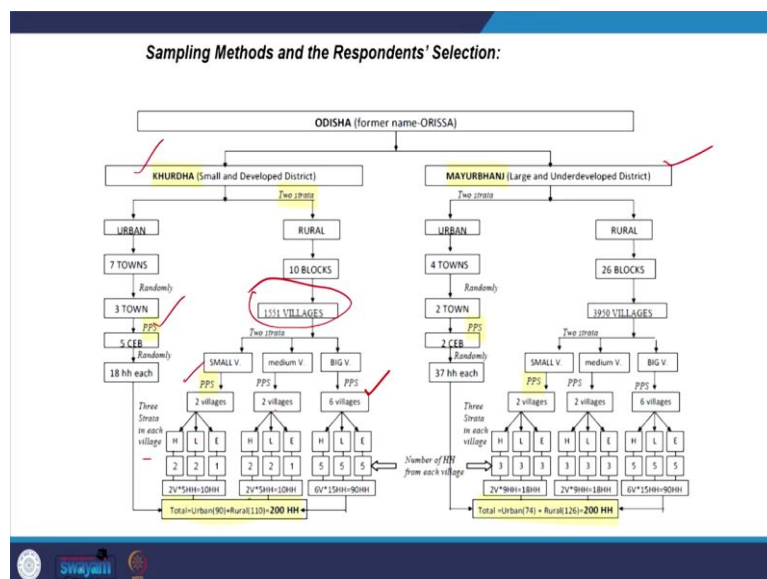
(Refer Slide Time: 33:10)

☐ NFHS-4 sampling and use of PPS

PPS is used in NFHS 4 to select PSUs (villages in rural area and census enumeration blocks in urban areas).

In NFHS 4 sampling PPS was also used and they considered villages in rural areas and census enumeration blocks in urban areas.

(Refer Slide Time: 33:23)

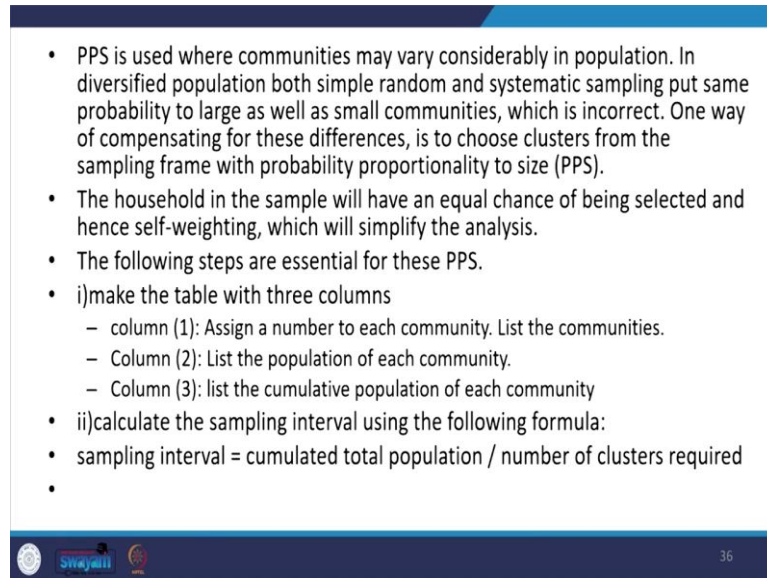


Like in my own study I did a survey in Odisha where two districts were selected for a study that is one is Kurdha and the second one is Mayurbhanj and a multistage sampling

method applied. But in the process finally, select the villages from out of so many villages and their population.

Finally, the village selection was made through PPS. Here also the census and enumeration blocks were collected through PPS techniques.

(Refer Slide Time: 34:15)



- PPS is used where communities may vary considerably in population. In diversified population both simple random and systematic sampling put same probability to large as well as small communities, which is incorrect. One way of compensating for these differences, is to choose clusters from the sampling frame with probability proportionality to size (PPS).
- The household in the sample will have an equal chance of being selected and hence self-weighting, which will simplify the analysis.
- The following steps are essential for these PPS.
- i) make the table with three columns
 - column (1): Assign a number to each community. List the communities.
 - Column (2): List the population of each community.
 - Column (3): list the cumulative population of each community
- ii) calculate the sampling interval using the following formula:
- $\text{sampling interval} = \text{cumulated total population} / \text{number of clusters required}$
-

PPS is used where communities may very considerable in population which I have already said. In diversified population both simple random sampling and systematic sampling put the same probability to large as well as small communities, which is incorrect.

One way of compensating for these differences is to choose the clusters from the sampling frame with probability proportional to size. The household in the sample will have an equal chance of being selected and hence self-weighted, which will simplify the analysis. I think in this case I will suggest that you should follow the NFHS report.

And it has guided you on how you can follow the PPS techniques. Some of the steps I have already guided you. You assign a number to each of the community then you list those communities by their population. In the third column, you can make them in cumulative population.

Then the next aspect is to calculate the sampling interval. There must be a systematic ordering followed to define to select the sample. The sampling interval basically is calculated based on the cumulative total population divided number of clusters.

(Refer Slide Time: 35:54)

- iii) select a random number: between 1 and the sampling interval
- iv) look back at the table and choose the village whose cumulative population exceeds this random number.
- v) Add the sampling interval to the random number
- vi) Choose the community whose cumulative population just exceeds this number. This becomes our second cluster or village. Accordingly select other required clusters.
- vii) Identify the location of each subsequent cluster by adding the sampling interval to the number which located the previous cluster. Stop when you have located as many clusters as you need.

$$\text{sampling interval} = \frac{\text{cumulated total population}}{\text{number of clusters required (big villages)}} = \frac{643428}{6} = 107238$$

And then select a random number between one and the first sample interval. Then you can look back at the table and choose the village whose cumulative population exceeds this particular number. And then we keep on adding the sample interval and the next cluster or next village or it can be selected. Similarly, we can follow and find out the number of the village to be selected.

The sampling interval is basically the cumulative population divided by the number of clusters required. This is from my own data and this was the total number and the required number of clusters was 6 from each of this unit we are supposed to select our samples.

So, you need not get confused with these numbers. This only gives you a rough idea about each cluster. The first cluster would be from one till this 107238.

Suppose you are selecting 1 number from the first cluster is around 100000 then you simply add the sampling interval that is 107238 which would be your second unit to be selected for the sample.

So, these are the way by which you can follow if you still have some difficulties in understanding the PPS technique. I will suggest that you will either follow the LASI report or you may follow the national family health survey report. This has given systematic guidance about probability proportional to size sampling.

And this is highly utilized these days especially in the multi-stage sampling procedure or in the large-scale survey data set. These are all for today I will look forward to your participation in the next class further on these issues with this I think I should close here.

Thank you.