IIHMR UNIVERSITY

www.iihmr.edu.in

Dr. J.P. Singh
Professor
IIHMR University, Jaipur

Faculty Development Program for IIHMR Group of Institutions

Statistical Survey Design: Basic Concepts

Date: May 15, 2021

Prof. J.P.Singh is a statistician. With an experience of more than 15 years in health sector, he has gained expertise in conducting health surveys and programme evaluations. He has been coordinating national and state level large scale surveys of repute, like National Family Health Survey (NFHS), District Level Household Survey (DLHS) and Facility survey under RCH programme. He has contributed in several programme evaluations, including evaluation of NACP, NLEP, etc. He has published several research papers in reputed journals. He has been teaching statistics in PGDHM programme, and facilitating management development programmes in statistics and survey research.

Statistical Survey Design
Basic Concepts

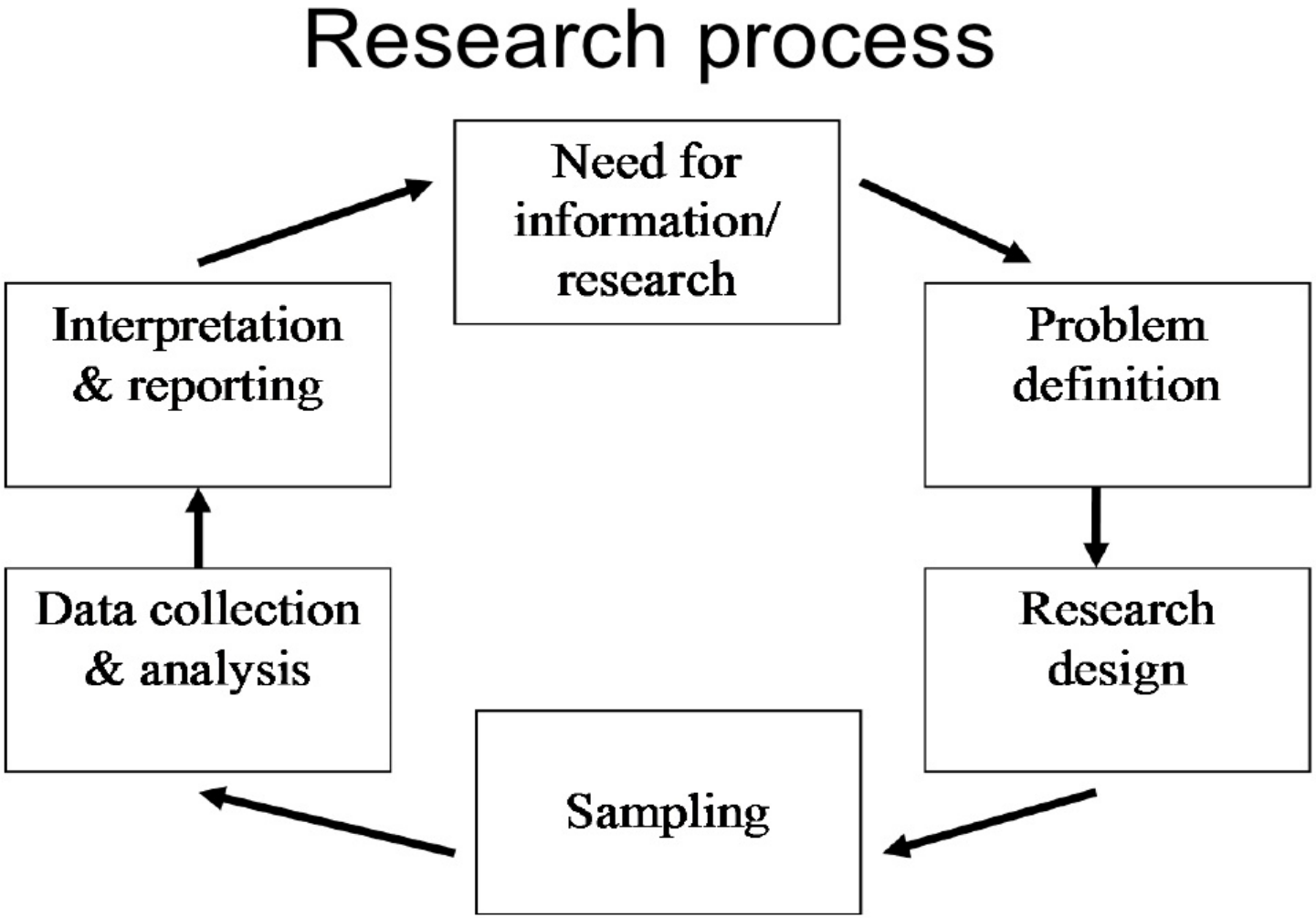
J.P.Singh

Presentation outline

- **Introduction**
  - Target population, study population, sampling unit, and study unit
  - Sample, sampling, and sampling frame
- **Bias and error**
- **Equal probability of selection method (EPSEM) / self-weighting design**
- **The need for sample weights**
- **Conclusion**

Note: The material used here is taken from mainly a book written by Prof. T.K.Roy (former Director, IIPS, Mumbai and JHU open-source material and from internet.

Introduction



Objective: Sample survey design

- Generalizability - To make inferences about a population parameter by observing a portion or sample from it
- Margin of error – It is natural to expect that this phenomenon of interpreting about a wider group (a population) on the basis of a sample from it will have some margin of error in the result
- Procedure of sample selection – How the samples are drawn or selected, while designing a sample survey, is important to ensure that the inference drawn is both valid and reliable
- Alternate design – The theory can also guide in choosing an alternative design so that the margin of error can be minimized

Target population

Target population (universe) - The entire group of people or objects to which the researcher wishes to generalize the study findings

Examples

- All institutionalized elderly with Alzheimer's
- All people with AIDS
- All low birth weight infants
- All school-age children with asthma
- All pregnant teens

Study population

Study population - The entire group of people or objects in a study which is considered for sample selection.

For example, to understand teenage pregnancy in an area, all girls aged 13-19 years residing in the area becomes the target population. However, while implementing the survey design, more often than not, only households that might have girls aged 13-19 years residing in it are considered for selection.

Sampling unit and study unit

Sampling unit - A sampling unit is an object that can be selected with known probability from a sampling frame.

Study unit – A study unit is a unit or a member of a study population which is a unit of analysis in a study.

Considering an example of sampling from a human population, it is often preferred to select households before selection of individuals – the study units in a population. In this case, the households, consisting of individuals, which facilitate the process of the sample selection are called sampling unit.

Sampling design

A sampling design is the process to accomplish the following tasks:

- Number of units to have in a sample (sample size)
- The procedure of selection of the of the required units (how to select the required units into the sample) : Different sampling technique

Sampling design

- The procedure of estimation of the required parameters (to transfer the information collected from the sample into meaningful measure to gauge the population parameter
- Provide an idea about the likely error that one would commit in making the inference from the sample to the target population

Sample size

The decision about sample size is made largely by considering (a) level (including sub-groups estimate) at which the estimates are to be produced, (b) a reasonable level of precision, and (c) availability of resources.

Probability sampling methods

- Also called random sampling
- Each and every unit in the study population is given a known and non-zero chance of being included in a sample
- The actual selection of a unit is a chance event and depends on its probability of selection

Probability sampling methods

- The group of units that will ultimately form a sample is not known in advance
- Once a design, in terms of sample size and the chance of selection of each unit are specified
- A variety of options or alternative samples will be possible
- It provide a basis for arriving at a valid estimation procedure and also an idea about the error in estimation

Frame

- A frame is the basis for drawing sampling units
- A frame is supposed to provide the list (detailed address to facilitate contacting a selected unit) of all the sampling units
- If there are N units in a population, the frame should have only N entries in the list giving contact details of each of the units

Frame

- The basic techniques, such as simple random sampling, stratified sampling and systematic sampling presume availability of such a list of elements if elements are being selected directly
- In cluster and multistage sampling, a list of all areal units comprising a population is required initially
- In multistage sampling, macro areal units are selected using a PPS sampling, which assumes an availability of size or at least an estimated size of each cluster

Frame

- Selection of elements in the final stage of a multistage design would require a frame consisting of elements in selected areas
- This is generally obtained through listing of dwellings or households in a population-based surveys
- Frame error results when the sampling frame is not an accurate and complete representation of the population of interest

Bias and error

Sampling error – As per sample survey design, there may be several alternative samples that could be selected. Each possible sample will have an error that is defined as the deviation between a sample estimate and the corresponding population parameter.

The difference between the sample mean and the population mean that is due to the random fluctuations in data that occur when the sample is selected

Bias and error

- The sampling theory helps get an estimate of the error for a design from a single sample
- The summary measure of error is denoted by mean square error (MSE)
- The MSE comprises of two components bias and error

Bias and error

Sampling bias

- Also called systematic bias or systematic variance
- The difference between sample data and population data that can be attributed to faulty sampling of the population
- Consequence of selecting subjects whose characteristics (scores) are different in some way from the population they are supposed to represent
- This usually occurs when randomization is not used

If $\bar{y}_i$ denotes the sample mean in the case of the ith alternative sample with $\bar{y}$ as the mean of all the sample means in a design, then

$$\text{MSE}(\bar{y}) = \frac{\sum_{i=1}^{K} (\bar{y}_i - \bar{Y})^2}{K} \tag{1.3}$$

$$= (\bar{y} - \bar{Y})^2 + \frac{\sum_{i=1}^{K} (\bar{y}_i - \bar{y})^2}{K} \tag{1.4}$$

The sampling theory shows that the mean of all possible sample means would be equal to the population mean or unbiased only when the design is equal probability of selection method (EPSEM); that is, it assigns equal chance of selection to each and every unit to be selected in a sample.

The second component of equation 1.4 gives an idea about the extent to which alternative sample means, in a design, vary from their mean, and is known as the sampling variance of mean and its square root as the standard error of the mean. Hence,

$$\text{MSE}(\bar{y}) = B^2 + \text{sampling variance of the mean} \qquad (1.5)$$

where B is the extent of the bias.

The square root of MSE is known as the standard error of the mean and for an unbiased estimate, MSE is equal to the sampling variance of the mean.

EXAMPLE 1.1

The daily wage of a population consisting of six persons is as follows:

The small population under consideration has six members A–F, whose daily wage is given below.

Person	A	B	C	D	E	F
Daily wage	50	60	1200	100	1400	70

For this small population, the MSE of sample mean is computed for a design where a sample of three individuals are selected one by one without replacement (i.e., once a unit is selected, it will not have another chance of selection) giving equal chance of selection to each units.

$$\text{The population mean } (\bar{Y}) = \frac{(50+60+\cdots+70)}{6} = \frac{2880}{6} = 480$$

TABLE 1.1 Estimates of average daily wage and deviation between sample mean and the population mean for the 20 alternative samples drawn from the population shown in example 1.1

S. No. of samples	Units in a sample	Sample mean ($\bar{y}$)	Deviation ($\bar{y} - \bar{Y}$)
1	ABC	436.67	−43.33
2	ABD	70.00	−410.00
3	ABE	503.33	23.33
4	ABF	60.00	−420.00
5	ACD	450.00	−30.00
6	ACE	883.33	403.33
7	ACF	440.00	−40.00
8	ADE	516.67	36.67
9	ADF	73.33	−406.67
10	AEF	506.67	26.67
11	BCD	453.33	−26.67
12	BCE	886.67	406.67
13	BCF	443.33	−36.67
14	BDE	520.00	40.00
15	BDF	76.67	−403.33
16	BEF	510.00	30.00
17	CDE	900.00	420.00
18	CDF	456.67	−23.33
19	CEF	890.00	410.00
20	DEF	523.33	43.33
Sum of deviation			0.0

Since the estimate is unbiased, the first component (equation 1.4) is zero, it will be

$$\text{MSE}(\bar{y}) = \frac{(-43.33)^2 + (-410.00)^2 + \cdots + (43.33)^2}{20} = 67953.33$$

Note that this will be equal to its sampling variance as the mean is an unbiased estimate of the population mean. Hence,

$$\text{SE}(\bar{y}) = \sqrt{\text{MSE}(\bar{y})} = 260.68$$

where SE $(\bar{y})$ is the standard error of mean in the design.

TABLE 1.2 Estimates of average daily wage and the deviation of sample mean from the population mean for the four alternative samples from the second design (one selected randomly from A, B, D and F and both C and E are included with certainty, see data in example 1.1).

S. No. sample	Units in sample	Sample mean	Deviation from population mean ($\bar{Y}$)
1	ACE	883.33	403.33
2	BCE	886.67	406.67
3	CDE	900.00	420.00
4	CEF	890.00	410.00

It can be seen that the deviations are not randomly distributed around the population mean, as in the previous design. The four alternative samples are all on the higher side. In this case,

$$\bar{y} = \frac{883.33 + \cdots + 890.00}{4} = 890.00$$

$$B = (890.00 - 480.00) = 410.00$$

It can be noticed that although the extent of bias is large, the alternative sample means are quite close to each other, and the variance from their mean is quite small, equal to 38.9. Therefore, applying equation 1.4 we get,

$$\text{MSE}(\bar{y}) = 410.00^2 + \frac{(883.33 - 890.00)^2 + \cdots + (890.00 - 890.00)^2}{4}$$

$$= 16813.89$$

$$\text{SE}(\bar{y}) = \sqrt{\text{MSE}(\bar{y})} = 410.05$$

Simple random sampling (SRS)

In SRS design, each and every sampling unit in a population receives equal chance of being included in a sample. This design is termed as EPSEM.

Let there be N units in a population from which n are to be selected for inclusion in a sample. The method of selection adopted is such that every unit, in a sample of n units, has a chance of selection equal to:

Probability of selection of a unit = n / N

Simple random sampling (SRS)

- The distinguishing feature of SRS is that, in addition to giving equal chance of occurrence to each unit, it also assigns all possible samples of size n an equal chance of selection.

As per this design in daily wage example :

Each person in the sample has equal chance (= 3/6 = ½) of selection.

Each sample of 3 persons will have equal chance (1/20) of selection.

Sampling process

- We employ some *randomization* process for sample selection so that there is no preferential treatment in selection which may introduce selectivity bias.
- Listing (sampling) Frame
- Random number table (from published table or computer generated)
- Selection of sample

Sampling process

Computer generated random numbers: (STATA output)

832645	573158	467460	838921	171721	152885
708009	285644	727733	343305	539264	907568
305761	995036	740619	054728	746425	713746
536405	504168	750032	367682	626278	855480
217862	782003	409660	155199	129514	484511
844905	296231	103727	053603	562252	219726
670523	707073	049209	830572	337034	716264
334920	023934	808901	740693	170372	095017
885588	384435	129958	303040	264636	858065
458268	058670	888935	064613	661404	411861
277649	076177	482951	876389	898190	927367
977683	759956	553916	983998	331578	981306

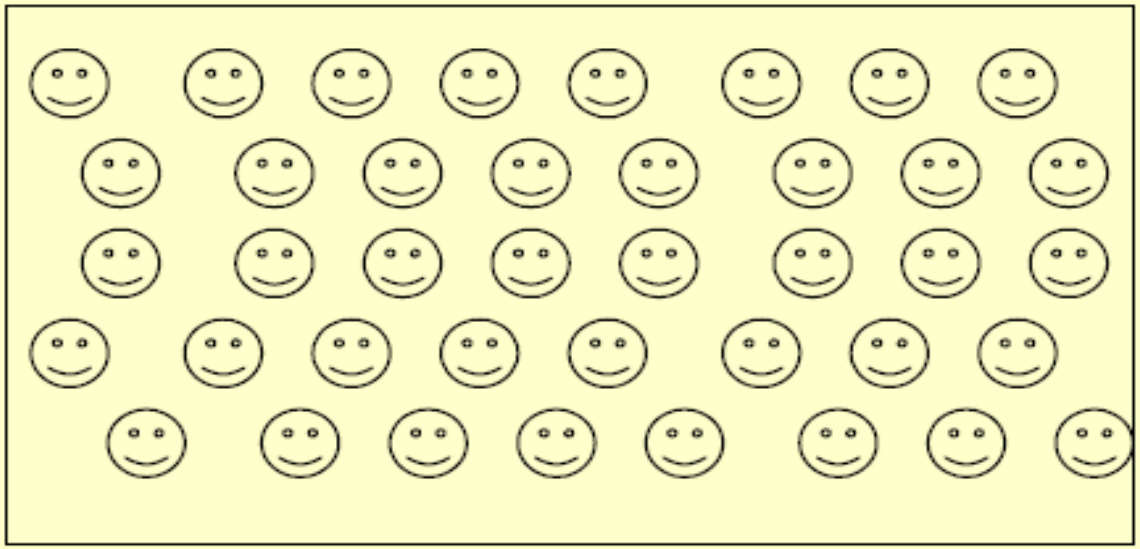
Systematic sampling

"…systematic sampling, either by itself or in combination with some other method, may be the most widely used method of sampling."

Levy and Lemeshow, 1999

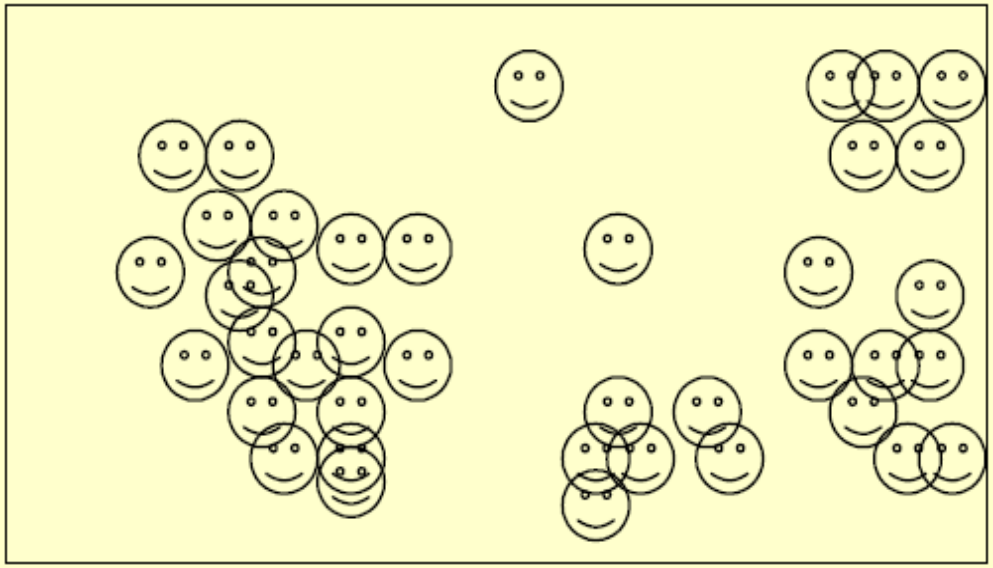
Systematic sampling

In simple random sampling we want that the samples should be distributed randomly.



Systematic sampling

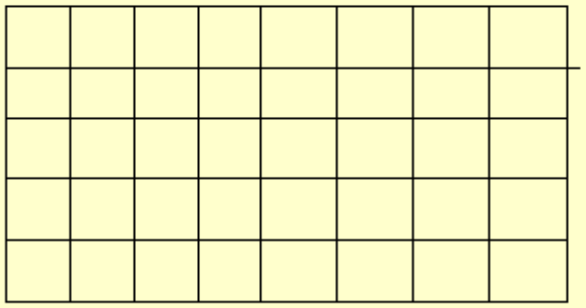
In reality the random selection may be like this:

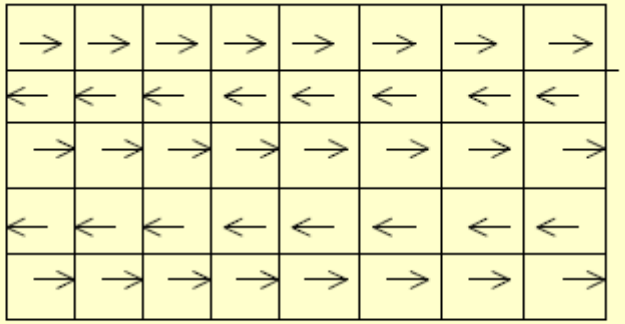


Systematic sampling

In systematic sampling we may force to select samples "evenly" from the list (sampling frame):

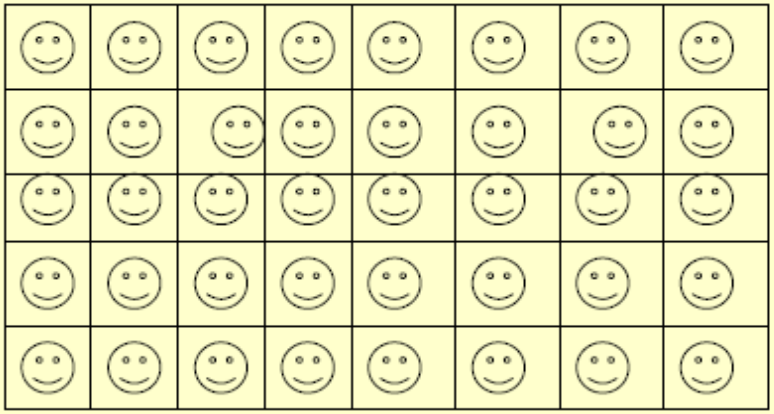
First, let us consider that we are dividing the list evenly into some "blocks".





Systematic sampling

Then, we select a sample element from each block.



Systematic sampling

- In systematic sampling, only the first unit is selected at random,
- The rest being selected according to a predetermined pattern.
- to select a systematic sample of *n* units,
  the first unit is selected with a random start *r*
  from 1 to k sample, where $k=N/n$ sample intervals,
  and after the selection of first sample, every k^{th} unit is included where $1 \leq r \leq k$.

Systematic sampling

An example:

Let N=100, n=10, then k=100/10.

Then the random start *r* is selected between 1 and 10 (say, r=7).

So, the sample will be selected from the population with serial indexes of:

7, 17,27,.....,97

i.e., r, r+k, r+2k,....., r+(n-1)k

[What could be done if k=N/n is not an integer?

Systematic sampling

Selection of systematic sampling when sampling interval (k) is not an integer

Consider, n=175 and N=1000. So, k=1000/175 = 5.71

One of the solution is to make k rounded to an integer, i.e., k=5 or k=6.

Now, if k=5, then n=1000/5=200; or,

If k=6, then n=1000/6 = 166.67 ~ 167.

Which n should be chosen?

Systematic sampling

Solution

if k=5 is considered, stop the selection of samples when n=175 achieved.

if k=6 is considered, treat the sampling frame as a circular list and continue the selection of samples from the beginning of the list after exhausting the list during the first cycle.

An alternative procedure is to keep k non-integer and continue the sample selection as follows:

Let us consider, k=5.71, and r=4.

So, the first sample is 4^{th} in the list. The second = (4+5.71) =9.71 ~9^{th} in the list, the third =(4+2*5.71) =15.42 ~ 15^{th} in the list, and so on. (The last sample is: 4+5.71*(175-1) = 997.54 ~ 997^{th} in the list).

Note that, k is switching between 5 and 6.

Probability proportional to size (PPS)

Communities may vary considerably in population. If simple random or systematic sampling is used, both large and small communities will have the same probability of being included, which is incorrect. One way of compensating for these differences is to choose clusters from the sampling

frame with probability proportional to size (PPS).

Procedure for PPS

To select the communities to be used in the survey you will need:

- a list of all communities in the district to be surveyed
- an estimate of population size (or the number of households) in each community
- the number of clusters you want to have in the district

Procedure for PPS

To sample the communities with PPS, complete the following seven steps:

Step1: Make a 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 - that is, the sum of the population of that community plus the populations of all the communities above it in the table.

Procedure for PPS

Step 2: **Calculate the sampling interval using the following formula:**

Sampling interval = cumulated total population ÷ number of clusters required

Using the example below, and supposing we need to select three clusters, this would give:

Sampling interval = 6,700 ÷ 3 = 2,233

Procedure for PPS

ILLUSTRATIVE TABLE:
Cumulating Community Populations

Village	Population size	Cumulative
1	1,000	1,000
2	400	1,400
3	200	1,600
4	300	1,900
5	1,200	3,100
6	1,000	4,100
7	1,600	5,700
8	200	5,900
9	350	6,250
10	450	6,700

Procedure for PPS

Step 3: **Select a random number which is equal to or less than the sampling interval**

For the above example, we need to choose a random number between 1 and 2,233. You can select this random number by using a table of random numbers. Let us suppose that the chosen number is 1,814.

Procedure for PPS

Step 4: **Look back at the table**

Locate the first cluster by finding the community whose cumulative population exceeds this random number. In the example, the first cluster would be located in village 4, where the cumulative population is 1,900.

Step 5: **Add the sampling interval to the random number**

In the example, 2,233 + 1,814 = 4,047.

Procedure for PPS

Step 6: **Choose the community whose cumulative population just exceeds this number**

The second cluster will be located in this community. In the example, the second cluster will be located in village 6.

Step 7: **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.

Stratified sampling

In stratified sampling the population is partitioned into groups, called strata, and sampling is performed separately within each stratum.

When?

- Population groups may have different values for the responses of interest
- If we want to improve our estimation for each group separately
- To ensure adequate sample size for each group

Stratified sampling

- stratum variables are mutually exclusive (non-over lapping), e.g., urban/rural areas, socio-economic categories, geographic regions, race, age, sex, etc.
- the population (elements) should be homogenous within-stratum, and
- the population (elements) should be heterogenous between the strata.

Equal allocation

- Divide the number of sample units *n* equally among the *K* strata.
- Formula: $nh = n/K$
- Example: n = 100; 4 strata; sample nh =100/4 = 25 in each stratum.
- May not be equal in each stratum. (what if you have 3 strata?)
- Need “weighted analysis” (disproportionate selection)

Proportional allocation

- Make the proportion of each stratum sampled identical to the proportion of the population.
- Formula: Let the sample fraction f= n/N.
- So, $n_h = fN_h = n(N_h/N) = nW_h$,
- Where $W_h = N_h/N$ is the stratum weight.
- Note, *f* is constant across strata, but W_h varies among strata.
- Self-weighted (equal proportion from each stratum)

Proportional allocation

Example:

- N = 1000
- n = 100
- *f = n/N=100/100=.1*
- *N_1 = 700 n_1 =fN_1 =0.1*700=70*
- *N_2 = 300 n_2 =fN_2 =0.1*300=30*

Cluster Sampling

In cluster sampling, cluster, i.e., a group of population elements, constitutes the sampling unit, instead of a single element of the population.

Need to consider the sampling order:

- Primary sampling units (PSU): clusters
- Secondary sampling units (SSU): households/individual elements

Selection process

- We may select the PSU's by using a specific *element* sampling techniques, such as simple random sampling, systematic sampling or by PPS sampling.

- We may select **all** SSU's for convenience or **few** by using a specific element sampling techniques (such as simple random sampling, systematic sampling or by PPS sampling).

Simple one-stage cluster sample:
List all the clusters in the population, and from the list, select the clusters – usually with simple random sampling (SRS) strategy. **All units** (elements) in the sampled clusters are selected for the survey.

Simple two-stage cluster sample:
List all the clusters in the population. First, select the clusters, usually by simple random sampling (SRS). The units (elements) in the selected clusters of the first-stage are then sampled in the second-stage, usually by simple random sampling (or often by systematic sampling).

Multi-stage sampling:
when sampling is done in more than one stage.
In practice, clusters are also stratified.

- Use as many clusters as feasible.
- Use smaller cluster size in terms of number of households/individuals selected in each cluster.
- Use a constant "take size" rather than a variable one (say 30 households from each cluster).

Design effect

It is informative to compare the variance in the case of a cluster design with that if units were selected directly (in one stage) using SRS. A ratio of the two variances is known as the 'design effect' (deff), that is,

$$\text{deff} = \frac{\text{sampling variance in cluster design of, say, } n \text{ units}}{\text{variance in the case of simple random sample of size } n} \qquad (3.11)$$

This ratio can also be obtained as (Kish, 1995, p. 161–162)

$$\text{deff} = 1 + (b - 1)\rho \qquad (3.12)$$

where ρ denotes the intra-cluster correlation and b is the cluster size (by cluster size we mean the number of units selected in a cluster or size of an ultimate cluster). The value of deff provides a comparison of relative precision of a cluster design to that of a simple random sample of similar size. It gives the extent of increase, if any, in the sampling variance because units are not selected in one stage using SRS but are drawn in two or more stages. The extent of increase depends on i) extent of intra-cluster correlation and ii) number of units (b) that are selected from each selected cluster.

Sample weights

- The purpose of weighting sample data is to improve ***representativeness*** of the sample in terms of:
  - **size**
  - **Distribution, and**
  - **characteristics** of the study population.
- By introducing sampling weight, the estimation is carried out in such a way that the estimates reflect the actual distribution/characteristics of the population.

The development of sampling weights

- It usually starts with the construction of the base weight for each sampled unit, to correct for their unequal probabilities of selection.
- In general, the base weight of a sampled unit is the reciprocal of its probability of selection into the sample.
- In mathematical notation, if a unit is included in the sample with probability pi, then its base weight, denoted by wi, is given by

  $$w_i = 1/ p_i.$$

- For example, a sampled unit selected with probability 1/50 represents 50 units in the population from which the sample was drawn.
- For multi-stage designs, the base weights must reflect the probabilities of selection at each stage.

Weighting for unequal probabilities of selection

8. An *epsem* sample of 5 households is selected from 250. One adult is selected at random in each sampled household. The monthly income (y_{ij}) and the level of education (z_{ij}= 1, if secondary or higher; 0 otherwise) of the *j-th* sampled adult in the *i-th* household are recorded. Let M_i denote the number of adults in household *i*. Then, the overall probability of selection of a sampled adult is given by:

$$p_{ij} = p_i \times p_{j(i)} = \frac{5}{250} \times \frac{1}{M_i} = \frac{1}{50} \times \frac{1}{M_i}$$

Therefore, the weight of a sampled adult is given by:

$$w_i = \frac{1}{p_{ij}} = 50 \times M_i$$

- *Example*

9. To illustrate the estimation procedure, let us assume a first-stage sample of 5 households with data obtained from the single sampled adult for each household as given in the table below:

Sampled Household	M_i	w_i	y_{ij}	z_{ij}	$w_i y_{ij}$	$w_i z_{ij}$	$w_i z_{ij} y_{ij}$
1	3	150	70	1	10,500	150	10,500
2	1	50	30	0	1,500	0	0
3	3	150	90	1	13,500	150	13,500
4	5	250	50	1	12,500	250	12,500
5	4	200	60	0	12,000	0	0
TOTAL	**16**	**800**	**300**	**3**	**50,000**	**550**	**36,500**

Estimates of various characteristics can then be obtained from the above table as follows:

a. The estimate of monthly income is

$$\bar{y}_w = \frac{\sum w_i y_{ij}}{\sum w_i} = \frac{50{,}000}{800} = 62.5$$

If weights were not used, this estimate would be 60 (=300/5)

b. The estimate of the proportion of people with secondary or higher education is

$$\bar{y}_w = \frac{\sum w_i z_{ij}}{\sum w_i} = \frac{550}{800} = 0.6875 \text{ or } 68.75\%$$

If weights are not used, this estimate would be 3/5 or 0.60 or 60%.

c. The estimate of the total number of people with secondary or higher education is

$$\hat{t} = \sum w_i z_{ij} = 550$$

d. The estimate of the mean monthly income of adults with secondary or higher education is

$$\bar{y}_w = \frac{\sum w_i z_{ij} y_{ij}}{\sum w_i z_{ij}} = \frac{36{,}500}{550} = 66.36$$

11. Note that for estimating totals, sampled elements need to be weighted by the reciprocal of their selection probabilities. For estimating means and proportions, the weights need only be proportional to the reciprocals of the selection probabilities. Thus, in the preceding example, the weights w_i's are proportional to Mi (w_i=50* M_i). If we use M_i as the weights, then the estimate of the proportion with secondary or higher education is

$$\hat{p} = \frac{\sum M_i z_{ij}}{\sum M_i} \frac{3\times1+1\times0+3\times1+5\times1+4\times0}{3+1+3+5+4} = \frac{11}{16} = 0.6875 \text{ or } 68.75\%,$$

as before. However, the estimate of the total number of adults with secondary or higher education is

$$\hat{p}_s = 50\sum M_i z_{ij} = 50\times11 = 550$$

The adjustment of sample weights for non-response

- **Step 1**: Apply the initial weights (for unequal selection probabilities and other adjustments, if applicable);
- **Step 2**: Partition the sample into subgroups and compute weighted response rates for each subgroup;
- **Step 3**: Use the reciprocal of the subgroup response rates for non-response adjustments; and
- **Step 4**: Calculate the non-response adjusted weight for the *i*-th unit as: $w_i = w_{1i} * w_{2i}$,

where w_{1i} is the initial weight and w_{2i} is the non-response adjustment weight. Note that the weighted non-response rate can be defined as the ration of the weighted number of interviews completed with eligible sampled cases to the weighted number of eligible sampled cases.

Example

A stratified multi-stage sample of 1000 households is selected from two regions (North and South) of a country. Households in the North are sampled at a rate of 1/100 and those in the south at a rate of 1/200. Response rates in urban areas are lower that those in rural areas. Let n_h denote the number of households sampled in stratum *h*, let r_h denote the number of eligible households that responded to the survey, and let th denote the number of responding households with access to primary health care. Then, the non-response adjusted weight for the households in stratum h is given by: wh = w1h*w2h, where w2h= nh/ rh. Assume that the stratum-level data are as given in the following table:

Stratum	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h r_h$	$w_h t_h$
North-Urban	100	80	70	100	1.25	125	10,000	8,750
North-Rural	300	120	100	100	2.50	250	30,000	25,000
South-Urban	200	170	150	200	1.18	236	40,120	35,400
South-Rural	400	360	180	200	1.11	222	79,920	39,960
Total	**1,000**	**730**	**500**				**160,040**	**109,110**

23. Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109{,}110}{160{,}040} = 0.682 \quad \text{or } 68.2\%$$

The estimated number of households with access to primary health care is

$$\hat{t} = \sum w_h t_h = 109{,}110 = 68.2\% \text{ of } 160{,}040$$

Note that the unweighted estimated proportion of households with access to primary health care, using only the respondent data is

$$\hat{p}_{uw} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0.685 \text{ or } 68.5\%,$$

and the estimated proportion using the initial weights without non-response adjustment is

$$\hat{p}_1 = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83{,}000}{126{,}000} = 0.659 \text{ or } 65.9\%.$$

The adjustment of sample weights for non-coverage

Example

In the preceding example, suppose that the number of households is known to be 45,025 in the North and 115,800 in the South. Suppose further that the weighted sample totals are respectively 40,000 and 120,040.

Step 1: Compute the post-stratification factors.

For the North region, we have: $w_{3h} = \frac{45{,}025}{40{,}000} = 1.126$; and

For the South region, we have: $w_{3h} = \frac{115{,}800}{120{,}040} = 0.965$.

Step 2: Compute final, adjusted weight: $w_f = w_h \times w_{3h}$

34. The numerical results are summarized in the following table:

Stratum	r_h	t_h	w_h	w_f	$w_f{*}r_h$	$w_f{*}t_h$
North-Urban	80	70	125	140.75	11,256	9,849
North-Rural	120	100	250	281.40	33,768	28,140
South-Urban	170	150	236	227.77	38,709	34,155
South-Rural	360	180	222	214.20	77,112	38,556
Total	**730**	**500**			**160,845**	**110,700**

Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p}_f = \frac{\sum w_f t_h}{\sum w_f r_h} = \frac{110{,}700}{160{,}845} = 0.688 \quad \text{or } 68.8\%$$

35. Note that with the weights adjusted by post-stratification, the weighted sample counts for the North and South regions are respectively 45,024 (11,256+33,768) and 115,821 (38,709+77,112), which closely match the control totals given above.

Weighting and its similarity with standardization

TABLE 2.5 Prevalence of tobacco use and number of persons by age in populations A and B.

Age group	Distributions in A		Distributions in B	
	Population (%)	Tobacco use rate (per 100)	Population (%)	Tobacco use rate (per 100)
20–34	2070 (34.5)	37.1	243 (40.5)	38.5
35–59	2946 (49.1)	45.6	217 (36.2)	47.2
60 and above	984 (16.4)	25.2	140 (23.3)	28.2
Total	6000 (100.0)		600 (100.0)	

The purpose is to compare tobacco use rates for the two populations. For this, it is customary to obtain a single index like the prevalence of tobacco use in the two. They are

$$\text{Prevalence of use in A} = \frac{2070 \times 37.1 + 2946 \times 45.6 + 984 \times 25.2}{6000} = 39.3 \qquad (2.32)$$

$$\text{Prevalence of use in B} = \frac{243 \times 38.5 + 217 \times 47.2 + 140 \times 28.2}{600} = 39.2 \qquad (2.33)$$

Weighting and its similarity with standardization

Tobacco use rate in A = 39.3 (as in equation 2.32)

$$\text{Tobacco use rate in B (using age distribution of A)} = \frac{2070x\ 38.5 + 2946x\ 47.2 + 984x\ 28.2}{6000} = 41.1 \qquad (2.34)$$

To estimate the overall prevalence of tobacco use in B (the sample estimate), the weight for the three age groups ($h = 1, 2, 3$) would be $\left(\frac{2070}{243}\right)$, $\left(\frac{2946}{217}\right)$ and $\left(\frac{984}{140}\right)$ and according to equation 2.27,

the overall prevalence

$$= \frac{\left(\frac{2070}{243}\right) \times \sum_{i=1}^{243} y_{1i} + \left(\frac{2946}{217}\right) \times \sum_{i=1}^{217} y_{2i} + \left(\frac{984}{140}\right) \times \sum_{i=1}^{140} y_{3i}}{6000}$$

$$= \frac{\left(\frac{2070}{243}\right) \times 243 \times 38.5 + \left(\frac{2946}{217}\right) \times 217 \times 47.2 + \left(\frac{984}{140}\right) \times 140 \times 28.2}{6000} \qquad (2.35)$$

$$= 41.1$$

Conclusion

- Although a probability sampling is not a must, it is a desirable choice
- It is must for undertaking a scientific investigation
- Purposive selection can lead to unknown biases that can be serious
- It provides an idea about the bias, if any, as well as error in an estimate

Conclusion

- As for as possible, an EPSEM design should be considered. It helps in obtaining an unbiased estimate of a population parameter
- There are situations where it is neither possible nor necessary to have an EPSEM design. In such situations, bias occurring due to unequal probability of selection of units can be compensated by using appropriate sample weights

Conclusion

- Error is generally less of a worry compared to bias in a design
- It is desirable particularly in large-scale sample surveys, to keep the design as close to EPSEM as possible, unless there are genuine reasons to believe that deviating from EPSEM would help in reducing MSE
- A sample design should endeavour to capture the heterogeneity or variation in the study variable

Thanks