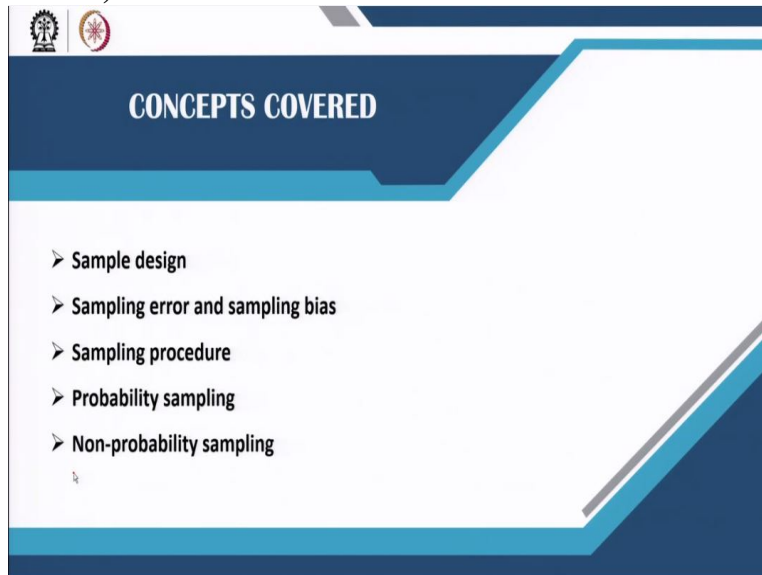


Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology - Kharagpur

Lecture-11
Sampling Theory-1

(Refer Slide Time: 00:32)



Welcome back. Module 3 will cover data collection and survey techniques and lecture 11 is on sampling theory part 1. So, the different concepts covered in this lecture are sample design, sampling error and sampling bias, sampling procedure, probability sampling and non-probability sampling.

(Refer Slide Time: 00:51)

Sample design

Samples should represent the population efficiently.

The entire population cannot be surveyed due to time and budget constraints.

Characteristic of Population (finite or infinite)

Sampling Unit (geographic location, special socio-economic group of people, ethnicity or religion, individual or household level survey)

Source List
Comprehensive list of all elements of any finite population is required for probability sampling. Non-probability based sampling technique otherwise.

Parameters of interest (Key determinant of the sampling unit, size and procedure)

Sampling Procedure

Size of Sample
(desired level of precision, acceptable confidence level, parameters of the population, the size of population, and the size of population variance).

Dr. Khanna

NPTEL

Background

Studies, like the one used in land use and transportation, require surveys. But due to time and monetary constraints an entire population cannot be surveyed. For this a subset of the population, called sample, is studied. Sample design involves choosing a sample of appropriate size that is representative of the entire population. The important components of sample design are enumerated below.

Sample design is based on the **characteristics of the population**. A population can be finite, like the population in an urban area or infinite, in case of sampling water. The **sampling unit** is based on geographical location, socio-economic status, and ethnicity. For carrying out probability sampling one may need to start off with a comprehensive and correct list of all the elements of the population. This is called the **source list**, like the list of all persons in an area. From this a sample can be randomly selected. In the absence of such a list, non-probability based techniques are used. The next important determinant for survey design is the **parameter of interest**. For example, if the survey concerns bicyclists in an area, then the parameter of interest is a person having a bicycle. So, people with a bicycle become part of the sampling unit only. **Sampling procedure** deals with the technique of surveying to be used. And **sample size** depends on the precision required (sampling error), acceptable confidence level, parameters of the population, the size of the population, and population variance.

(Refer Slide Time: 05:55)

Sampling error and sampling bias

Sampling error can be reduced with appropriate sample size determination.

Sampling bias is caused due to inappropriate selection of sampling technique, sampling frame etc.


- Sampling error creates impact on the variability of estimated parameters' average whereas sampling bias directly impacts on the value of that average.
- Sampling error is related with the precision of sample survey whereas sampling bias is related with accuracy of sample survey.

	ACCURATE	INACCURATE
PRECISE		
IMPRECISE		

● Actual avg. value ● Sample values

Accuracy (how close is an observation near the actual value)
Precision (how close are repeated observations near to observed values)

Difference between Accurate and Precise Solution



NPTEL Online Certification Courses
IIT Kharagpur

Sampling error and sampling bias

There are two types of error associated with survey sampling such as sampling error and sampling bias. Sampling error crops up because instead of studying the entire population, a small sample is studied. Every time a new sample is drawn from the population, the sampling error varies. Also, the larger the size of the sample, lower is the sampling error, generally.

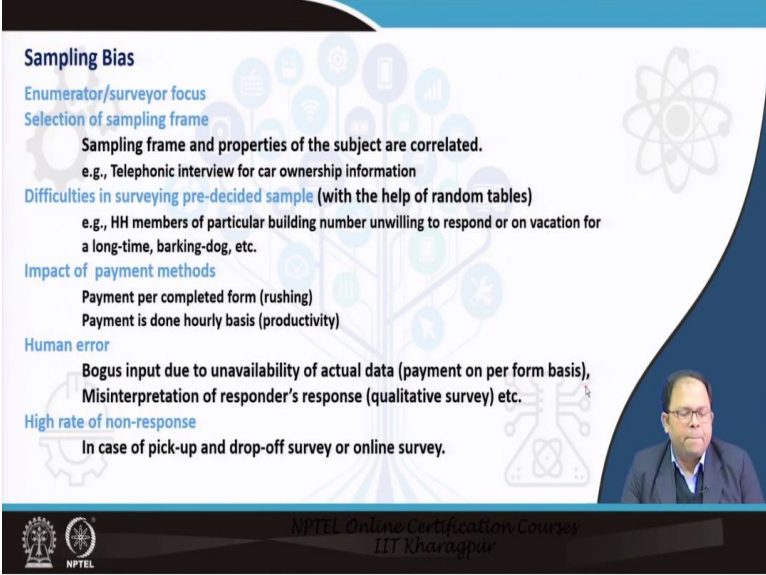
Sampling bias is caused due to inappropriate selection of sampling technique, sampling frame etc. Sampling frame is the method of collecting the samples. For example, somebody is conducting the telephonic interviews to collect samples. So, that is the sampling frame. So, sampling bias is caused by incorrect sample collection technique and has nothing to do with the sample size. Sampling error creates impact on the variability of the estimated parameters average, whereas sampling bias directly impacts the value of the average. Sampling bias results in a sampling mean very different from the population mean. Sampling error, on the other hand, results in variability of the estimated parameters.

So, an average of all these draws for the mean value will show the mean to be near to the population mean, but the variability would be high, i.e. the standard error or the standard deviation of this observation would be high. So, sampling error is related with precision and precision means how close are repeated observations near to the observed values.

A sampling bias is related with accuracy of sample survey. So, when sampling bias is there, then the estimates are not accurate. For example, in this case, one can see even though the readings are precise, that estimate is wrong.

If a study is so designed, that it is studying only a certain type of people in the entire population and not all the different kinds of people, it will lead to an inaccurate result. This is a sampling bias. This is a case of both inaccuracy and imprecision.

(Refer Slide Time: 11:04)



Sampling Bias

- Enumerator/surveyor focus**
- Selection of sampling frame**
 - Sampling frame and properties of the subject are correlated.
e.g., Telephonic interview for car ownership information
- Difficulties in surveying pre-decided sample (with the help of random tables)**
 - e.g., HH members of particular building number unwilling to respond or on vacation for a long-time, barking-dog, etc.
- Impact of payment methods**
 - Payment per completed form (rushing)
 - Payment is done hourly basis (productivity)
- Human error**
 - Bogus input due to unavailability of actual data (payment on per form basis),
 - Misinterpretation of responder's response (qualitative survey) etc.
- High rate of non-response**
 - In case of pick-up and drop-off survey or online survey.

NPTEL Online Certification Course
IIT Kharyapur

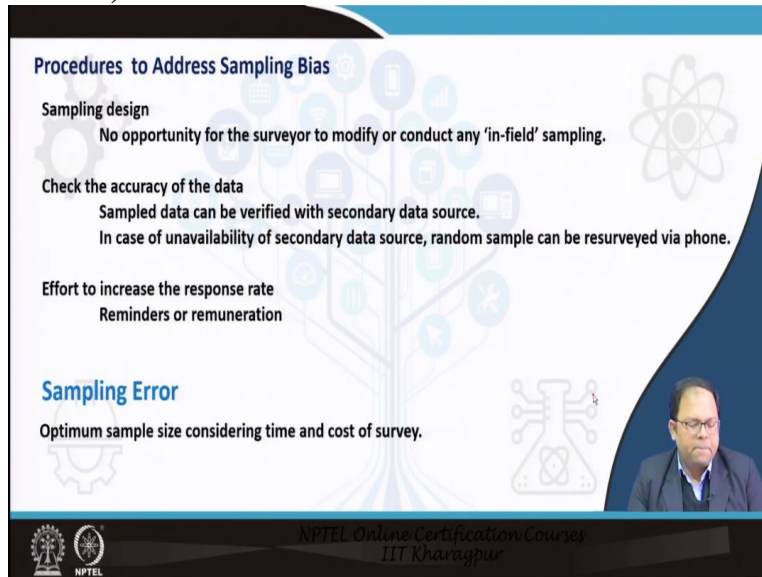
Sources of sampling bias

Sampling bias can be a result of enumerator or surveyor focus. That means, a surveyor, may choose to focus on certain groups and he/she misses out on the other people. So, that leads to some amount of bias from the enumerator's point of view.

Depending on the sampling frame, there can also be a bias. For example, if a telephonic interview is conducted for car ownership, it results in a bias. This is because telephone ownership and car ownership are correlated.

Sometimes a pre-decided sample cannot be surveyed or the enumerator is unable to fill up the entire questionnaire due to unavoidable reasons. When enumerators are engaged on an hourly basis, quality of data collected can go down. There is also a possibility of misinterpreting the response on the part of the enumerator. So, these are the different forms of sampling bias that can occur.

(Refer Slide Time: 14:14)



Procedures to Address Sampling Bias

- Sampling design**
No opportunity for the surveyor to modify or conduct any 'in-field' sampling.
- Check the accuracy of the data**
Sampled data can be verified with secondary data source.
In case of unavailability of secondary data source, random sample can be resurveyed via phone.
- Effort to increase the response rate**
Reminders or remuneration

Sampling Error
Optimum sample size considering time and cost of survey.

NPTEL Online Certification Course
IIT Kharagpur

Removing Sample bias

To remove sample bias, the survey questionnaire should be unambiguous. This will weed out surveyor discretion. And then the accuracy of the data should be verified with a secondary source. If secondary sources are not available then for part of the survey can be rechecked via telephonic interviews. Also incentives can be offered to the respondents to improve response rate. So, these are the different ways we can address sample bias.


An optimum sample size can reduce both sample error and the time and cost of the survey. This will be taken up at the next lecture.

(Refer Slide Time: 15:42)

Sampling Procedure

Probability sampling and non-probability sampling.
Restricted sampling and unrestricted sampling.

Sampling procedures	
Probability sampling	Non-probability sampling
Simple Random Sampling ✓	Convenience sampling ✓
Cluster Sampling ✓	Convenient sampling ✓
Systematic Sampling ✓	Judgement sampling ✓
Stratified Sampling ✓	Quota sampling ✓
	Snowball ✓
Unrestricted Sampling ○	Restricted Sampling ○



NPTEL Online Certification Courses
IIT Kharagpur

Sampling procedure

One way of categorizing sampling procedure can be in terms of probability and non-probability sampling. In probability sampling, unlike non-probability sampling, every member has an equal chance to be drawn into a sample from a particular population. Another categorization can be in terms of restricted sampling and unrestricted sampling. In restricted sampling there are certain rules or certain restrictions that are first put in and then a survey is conducted.

The table shows different types of probability and non-probability sampling techniques. The restricted ones are marked in gray, whereas non-restricted ones are marked in yellow. Under probability sampling there are different kinds of probability sampling like simple random sampling, cluster sampling, systematic sampling, stratified sampling and so on. Whereas in nonprobability sampling, there is convenience sampling, convenient sampling, judgment sampling, quota sampling and snowball sampling.

(Refer Slide Time: 17:44)

Probability Sampling

Every element of the population has equal and independent chance of getting selected as a sample ('random sampling' or 'chance sampling')

Simple Random Sampling

Samples are collected from the whole population randomly

Advantages :

- Each element has equal and independent chance of being included in the sample
- Simple and less sampling-error.

Disadvantages :

- Costly, thus demands larger sample size
- Full list of population is required before the sampling

Selection of sample: With replacement and Without replacement

In transportation planning,
Random sampling without replacement (diversity of the population)

NPTEL Online Certification Courses
18:19
T. Khanapur

Probability sampling

In probability sampling every element of the population has equal and independent chance of getting selected as a sample and it is either called random sampling or chance sampling. Random sampling or chance sampling is the most common form of sampling that people undertake.

Probability sampling has got certain advantages. There each element has an equal and independent chance of being included in the sample. Thus, the chance of sampling bias and sampling error is less and it is a simple survey. Whereas, the disadvantages are, that it requires a much larger sample size. So, it is also costly. However, the main problem is that, the full list of population is required before we start the sampling. Simple random sampling can also be done with replacement and without replacement. When the sampling is done with replacement, a person has the chance of getting drawn again in the next draw also. When sampling is done without replacement a person is not getting another chance to be drawn in the sample. In transportation planning, most of the random sampling is done without replacement to improve the heterogeneity or the diversity of the population.

(Refer Slide Time: 19:57)

Simple Random Sampling

Example

Household survey:
6 samples from 200 Households
(HH numbers 1-200 known beforehand)

Random number table:
First point of selection (any row and column number).

As the total population is of 3 digit (200) we will select only first 3 digits.

First sample number = 723. Rejected since > 200.


Next, we may go either row-wise or column-wise.

Random Number Table (Rand Corporation, 1955)

13642	30902	81322	20531	02100	88134	64621	70205	64763	03144
43995	46944	77810	11641	42148	34915	75061	13446	25261	89185
09564	48949	49999	39953	31918	42973	51021	43968	41078	86126
43174	32128	47137	35351	27084	02140	40387	39847	50968	96773
43753	31310	10430	50765	42393	41377	68164	29911	31391	72445
83563	15662	23336	48192	84284	38754	84753	34053	84502	29115
33607	71245	41628	44627	23077	99613	42603	03884	09751	48394
33320	10465	07107	41402	25424	89515	40911	24740	45407	45739
82415	80984	79750	87971	60322	34415	20882	12909	99476	40548
02612	26346	64418	08213	08181	57970	49520	79364	37474	15712
06936	37279	12813	71113	83023	40683	74665	12178	10761	58362
84861	64541	10454	30261	85111	40089	97019	21257	14899	47920
66164	88441	05193	06784	34714	64740	81087	83976	81301	72018
48602	94300	34440	42613	00711	66880	52584	96270	12122	64399
78401	77919	74678	76417	47983	74611	64678	79687	91821	21881
83111	13803	10343	78827	46961	80188	70844	29117	27921	16440
63841	84441	86677	91707	40384	29671	64084	77064	11611	81111
19120	02441	77784	33446	41304	30067	31584	70460	66664	79488
33717	03881	30914	43306	70100	86997	46561	79018	24271	25196
99789	06661	43946	43070	74928	40601	47790	46378	46144	59665
36140	38190	77705	28992	12126	34281	86227	64116	19426	00080
05026	46410	44916	79643	02830	77410	50021	37468	13845	77110
83962	29758	02795	04508	21289	04684	47961	23821	71451	98183
35114	46958	84446	21018	31177	44481	46814	48171	81614	47161
42122	40441	02140	70159	44168	38213	46889	20538	39983	67645
43626	40200	31483	34480	70280	24218	24596	04744	49106	30030
97161	04441	44488	24102	02088	77021	64000	15062	14401	64641
48175	44270	32012	09561	21661	10867	77611	70073	45142	23811
11370	71210	04616	71521	13412	46288	05012	20113	11489	48461
04487	24851	43879	07613	25403	17180	10880	66881	7196	10818
99448	07411	30447	80711	01765	57480	40885	57616	30051	37381
02420	76210	71449	14465	19017	14401	41448	79911	01011	39481
74813	40171	07090	79117	30089	97915	18305	42613	81251	79008
40462	76068	66756	30851	02102	50489	01481	31000	14647	47994
13813	30391	10811	90172	89149	66499	88162	71789	19964	02611
88448	02012	81918	20740	71176	84989	05514	12001	42947	20114
71018	70811	24118	19016	02081	81050	46444	99916	47990	11911
14995	38311	10018	37131	89189	76207	71222	99815	21204	20941
11892	21011	40910	30911	40184	00111	11201	94910	12044	84910
66009	28809	11839	60576	89161	40018	14200	47449	83837	52332
43101	04210	41016	70261	11067	14327	02044	20101	17149	00441
14010	40000	01011	76437	89145	84415	64763	94729	17075	02061
11844	08911	00110	81121	47278	90788	21842	84271	47914	63670
03843	82017	09112	09112	21286	51481	38113	13460	43064	43071

Final Samples:
162, 187, 161, 159, 30, and 5.

Random Number Generators



NPTEL Online Certification Courses
Kharagpur

Simple random sampling

This is an example of how random sampling is done. In this case, a random number table prepared by Rand Corporation is used. One can also use random number generators in the computer, like in Excel. This kind of table can be carried in the field and can be used to decide whom to survey. Let us suppose there is a requirement to collect 6 samples from 200 households. Households numbered 1 to 200 are known. Starting from any row or column a number is selected. Since, the households are numbered between 1 and 200 (i.e. a three digit number) the first three digits of the selected cells are looked into. The first number is 723 which are more than 200 and this is not considered. The next numbers are 380 and then 472, which are similarly rejected. The next cell is 162, and the household numbered 162 is selected. This process is repeated till six households are randomly identified.

(Refer Slide Time: 23:04)

Stratified Random Sampling (Restricted sampling)

Elements within each stratum are homogenous in nature and between strata heterogeneous

Simple random sample from a stratified population
Sample size from all the strata is same.

Stratified random sample from a stratified population
Sampling fraction instead of collecting same sample size from each stratum.

Advantages :
Sample size in each stratum can be controlled for highly skewed data.

Disadvantages :
Detailed information of the population is required.
Stratification is challenging in case of classification considering multiple characteristics.

NPTEL Online Certification Course
IIT Kharagpur

Stratified random sampling

The next procedure is stratified random sampling, which is a restricted sampling procedure. It is restricted because the population or sample is stratified. There are two types of stratified random sampling. The first type is simple random sampling from a stratified population and the other is stratified random sampling from a stratified population.

In stratified random sampling the entire population is divided into different groups having different characteristics. Thus, elements within each stratum are homogeneous in nature and there is heterogeneity between strata. And a survey of each group is carried out, with equal number of samples from each stratum. But if the strata have unequal populations this can result in sampling errors. In such a case, stratified random sample from a stratified population takes care of the problem. In this case the number of samples collected from each stratum is equivalent to its share in the population. Thus, sampling fraction is considered in stratified random sampling from a stratified population, which helps when the data is highly skewed. Stratification becomes challenging when multiple characteristics are used for stratification of the population instead of one.

(Refer Slide Time: 26:17)

Multi-stage Sampling

No prior information is required. Applicable when requirement for city/national level data collection.
e.g., National level travel diary survey in urban areas, India.



- Divide the population at state level and consider each state as total population in the next step
- Divide the states into districts and select sample districts
- Identify the urban areas in each sampled districts and collect sample urban areas
- Divide the urban areas in Travel Analysis Zones (TAZ) and enumerate each household within each TAZ
- Select sample households from TAZs based on simple random sampling technique

Advantages :

- Time requirement is less.
- No need to assign unique id to entire population.
- Different sampling technique can be adopted at each stage.

Disadvantages :

- Accuracy suffers.

Multi-stage sampling

In multistage sampling no prior information is required. But this is applicable in case when the data collection requirement is at a very high level, like a travel diary survey at the national level. It is very difficult to survey a sample from the entire population for a huge sample size. For this the population at the state levels is considered which are further divided into districts. A few sample districts are considered. From the selected districts a few urban areas are sampled. And then within each of the sampled areas, one can determine the TAZ (travel analysis zones), where sampling can be done based on simple random sampling technique. As the population is broken into so many levels, the technique suffers from a lack of accuracy.

(Refer Slide Time: 28:46)

Cluster Sampling



- The total population is divided into **several groups/cluster** (based on certain criteria, primarily, **geographical location/area sampling**) and then clusters are **selected randomly**.
- The population should be as **heterogeneous within cluster** and represent the total population.
- Clusters should be homogeneous.
- All the elements** within the selected clusters are treated as selected samples.
- In **two stage cluster sampling** a random sampling technique can be applied on the elements of each cluster.

Advantages: Cost is lower than simple random sampling.
Disadvantages: High sampling error which increases with similarity within the cluster.

Systematic Sampling

Every i^{th} element after the randomly selected starting point is selected as sample.

Advantages: Easy to design by any enumerator and less expensive.
Disadvantages: Not exactly as random as simple random sampling.

Cluster sampling

In this case the total population is divided into several groups or clusters based on some criteria, and then some of these clusters are selected randomly. The population within a cluster displays the heterogeneity of the entire population. That is why only a few clusters can be selected for studying in detail. All the elements within the selected clusters are then treated as selected samples and one can determine the different parameters based on this particular sample.

Systematic sampling

In this case, every element after a randomly selected starting point is selected as sample. For example, if 5 households are to be selected from a list of 200 households and a starting household is randomly selected at 30. To this number we add 40 ($200/5 = 40$) to obtain the subsequent samples, i.e. 70, 110, 150, and 190. This technique is easy to design and conduct and less expensive, though the degree of randomness is reduced.

(Refer Slide Time: 31:31)

Non-probability Sampling

These surveys are useful for **qualitative studies** and **exploratory studies** (i.e. pilot survey), suitable for identification of subjects which can be further explored or for broad understanding of the population.

Advantages: Requires less time and cost
Disadvantages: Results in lot of bias.

Convenience sampling

Only **available responders** are surveyed.
e.g., Random people whoever is available / willing to respond are surveyed in a market area or park.

Advantages: Simple and fast procedure which saves money, time and effort.
Disadvantages: Creditability and rationality associated with this method is low.

NPTEL Online Certification Courses
IIT Kharagpur

Non-probability sampling

These surveys are mostly useful for qualitative studies and exploratory studies like a pilot survey (a survey before the actual survey to study the characteristics of the population). This helps in developing a broad understanding (within the limits of time and cost) of the population so that one can actually design the sampling procedures. As samples are not randomly selected, there is a high chance of bias.

The first type of non-probability sampling is convenience sampling, where only available responders are surveyed. For example, if one wants to find out the people who use both Para

transit and bus, the enumerator can go directly to a Para-transit stop and bus stop and randomly survey people instead of going to a household, where many people do not even use Para transit or bus. While the credibility and rationality associated with this method is of course lower, the procedure is simple, fast and cost effective.

(Refer Slide Time: 34:01)

Consecutive Sampling
Convenient available responder/ group of responder from the population is surveyed and studied over a period of time before moving to the second group of samples.

Quota sampling
Population is first divided into different groups/strata and then interviewers obtain responses from each strata/quota.
Sampling is done on the field and extra care is required to make the survey random.
Disadvantages: Preferential bias in parameter estimation.
Certain parameters are over represented and others not since surveyors select easily available responders.

NPTEL Online Certification Courses
IIT Kharagpur

Consecutive sampling

Consecutive sampling is almost similar to convenient sampling. But here, available responders or groups of responders are studied over a period of time before the next group of samples is taken up. So certain characteristics are determined with respect to the first sample before moving onto the next set of samples, where certain modifications can be adopted based on the study of the first group of samples.

Quota sampling

In quota sampling, the population is divided into different groups and samples are taken from each group to meet a quota. The sampling is done on the field and extra care is required to make the survey random.

There may be a preferential bias in parameter estimation, where the enumerator may prefer certain groups or certain service or certain household type and ignore other kinds of households. With this kind of bias creeping in certain parameters are over represented and others are not.

(Refer Slide Time: 35:28)

Judgmental or Purposive sampling

Responders are selected based on researchers' preconceived notion whom they found suitable and useful. This method is useful for expert opinion survey.

Snowball sampling

Enumerators first approach available responder who is suitable for the subject. Responder is then asked for contacts of the other potential responder. Useful when target group is very difficult to find. e.g., Searching for persons affected by particular diseases.

NPTEL Online Certification Courses
IIT Kharagpur

Judgmental or purposive sampling

Here responders are selected based on the researcher's preconceived notion of suitability and usefulness. For example, in expert opinion survey, the surveyor determines who is an expert in their particular field and enumerates his/her opinion.

Snowball sampling

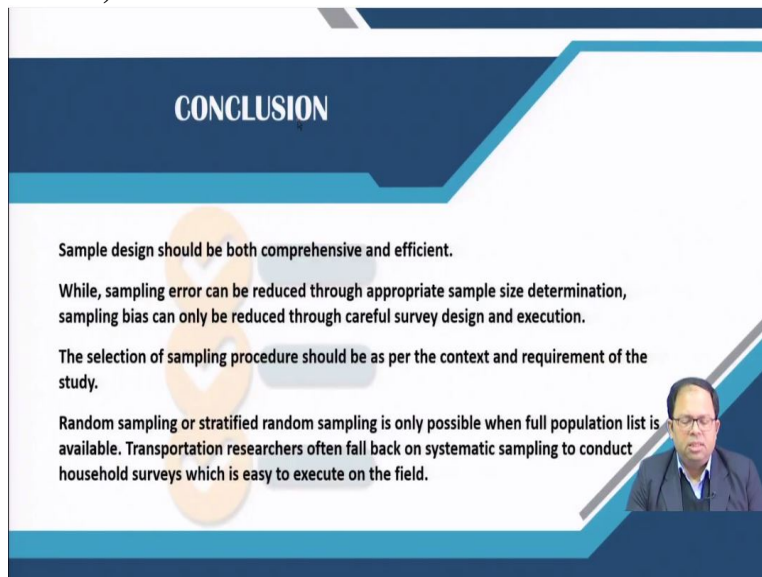
In snowball sampling enumerators approach available responders suitable for the particular study and then the responder is asked for contacts of other potential responders. This technique is very useful when the target group is very difficult to find. For example, people suffering from a particular disease can be traced through doctor's appointment or various support groups.

(Refer Slide Time: 36:51)

REFERENCES

- Abraham, V. M., Walpole, R. E., & Myers, R. H. (1979). Probability and Statistics for Engineers and Scientists. The Mathematical Gazette. <https://doi.org/10.2307/3616039>
- Bryman, A. (2003). Quantity and Quality in Social Research. In Quantity and Quality in Social Research. <https://doi.org/10.4324/9780203410028>
- Bryman, A. (2012). Social research methods Bryman. OXFORD University Press. <https://doi.org/10.1017/CBO9781107415324.004>
- Richardson, A. J., Ampt, E. S., & Meyburg, A. H. (1995). Survey Methods for Transport Planning.

(Refer Slide Time: 37:02)



CONCLUSION

- Sample design should be both comprehensive and efficient.
- While, sampling error can be reduced through appropriate sample size determination, sampling bias can only be reduced through careful survey design and execution.
- The selection of sampling procedure should be as per the context and requirement of the study.
- Random sampling or stratified random sampling is only possible when full population list is available. Transportation researchers often fall back on systematic sampling to conduct household surveys which is easy to execute on the field.

Small video inset of a man speaking in the bottom right corner of the slide.

Conclusion

Sample design should be both comprehensive and efficient. While sampling error can be reduced through appropriate sample size determination, sampling bias can only be reduced through careful survey design and execution. The selection of sampling procedures should be as per the context and the requirement of the study. Random sampling or stratified random sampling is only possible when full population list is available. Transportation researchers often fall back on systematic sampling to conduct household survey which is easy to execute on the field.