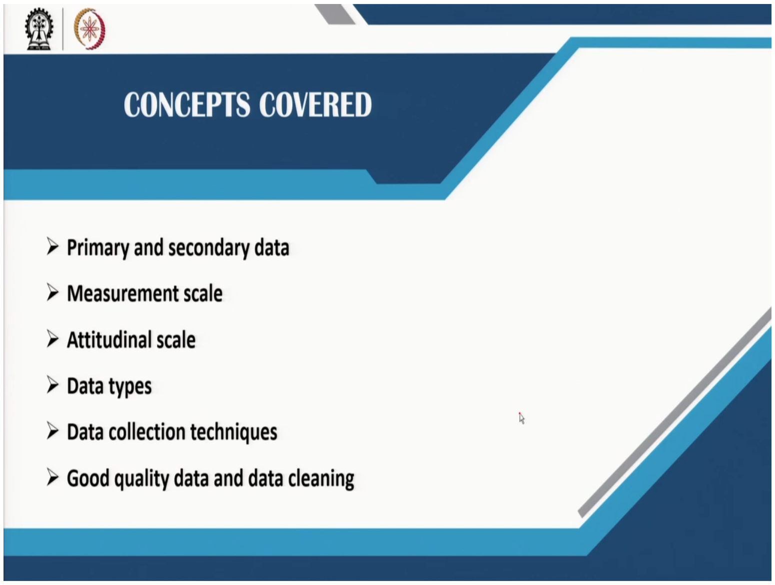
Urban Landuse and Transportation Planning

Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology – Kharagpur

Lecture-13
Data and Surveys

(Refer Slide Time: 00:27)

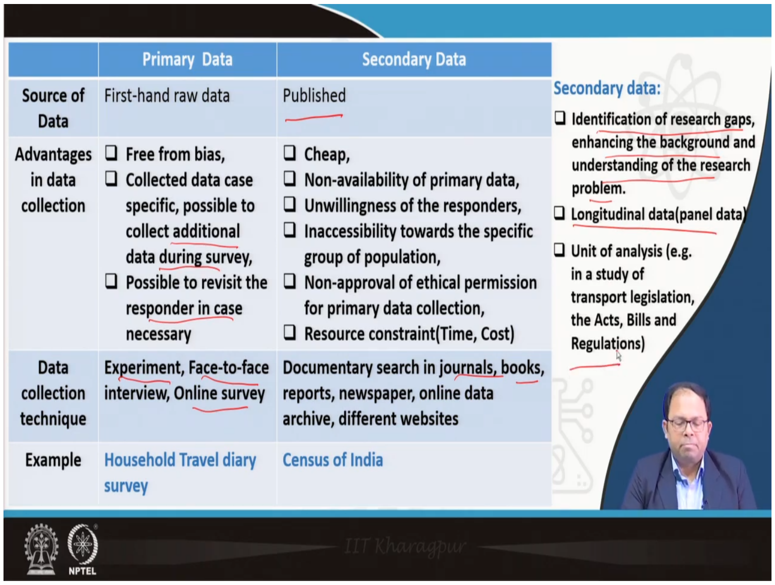


CONCEPTS COVERED

- Primary and secondary data
- Measurement scale
- Attitudinal scale
- Data types
- Data collection techniques
- Good quality data and data cleaning

Welcome back. Lecture 13 will cover data and surveys. The different concepts covered in this lecture are primary and secondary data, measurement scales, attitudinal scales, data types, data collection techniques and good quality data and data cleaning.

(Refer Slide Time: 00:46)



	Primary Data	Secondary Data
Source of Data	First-hand raw data	Published
Advantages in data collection	❑ Free from bias, ❑ Collected data case specific, possible to collect additional data during survey, ❑ Possible to revisit the responder in case necessary	❑ Cheap, ❑ Non-availability of primary data, ❑ Unwillingness of the responders, ❑ Inaccessibility towards the specific group of population, ❑ Non-approval of ethical permission for primary data collection, ❑ Resource constraint(Time, Cost)
Data collection technique	Experiment, Face-to-face interview, Online survey	Documentary search in journals, books, reports, newspaper, online data archive, different websites
Example	Household Travel diary survey	Census of India

Secondary data:

- ❑ Identification of research gaps, enhancing the background and understanding of the research problem.
- ❑ Longitudinal data(panel data)
- ❑ Unit of analysis (e.g. in a study of transport legislation, the Acts, Bills and Regulations)

Primary and secondary data

Depending on the source, data can be both primary and secondary. Primary data is raw data, collected first-hand, directly through experiments or through face to face interviews or even through online surveys. One good example is of course the household travel diary survey. This kind of data is relatively free from bias. Additionally, the surveyor can also revisit the responder in case there is any requirement, for example, when some data is missing.

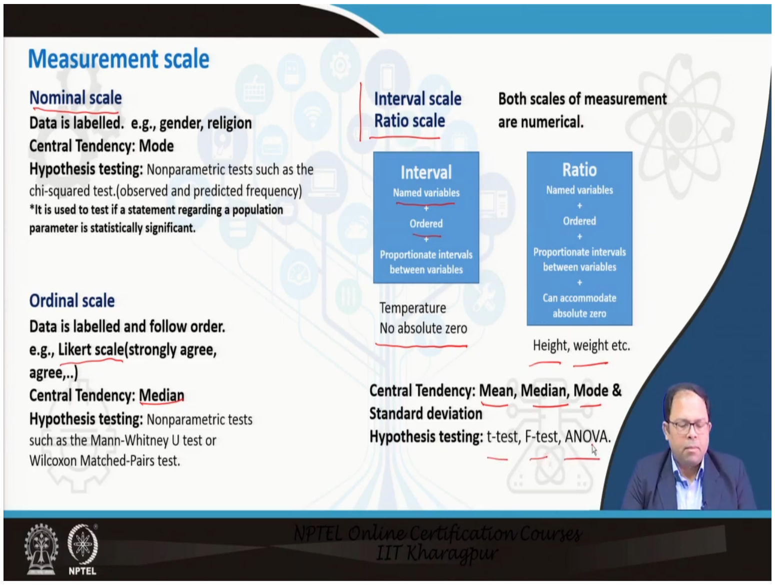
Secondary sources of data are already published and they can be obtained from documentary search of journals, books, reports, newspaper, online data archive, and different websites and so on. One good example is the Census of India data set. From there one can get data for different urban areas, rural areas, household data, and all the other data that is collected during census survey.

Secondary data is cheaper compared to primary data because investment in a new survey is not required. Often secondary data is old, making primary data collection inevitable. But primary data collection involves many hassles. Like responders are unwilling to share data or sometimes the situation is such that surveyors do not have access to a particular population or are denied permission because of ethical reasons or resource constraints. In such a situation secondary data becomes useful.

In most cases primary data and secondary data are both used. Secondary data is usually required to identify the research gaps, enhancing the background and understanding of the research problem. The gaps identified can be covered using a primary survey.

One can get longitudinal data or panel data from previous years which can be used for projections which are in turn matched with the existing data to see if the projection matches. Also, secondary data can itself act as a unit of analysis. For example, a study on transport legislation and regulation itself becomes data.

(Refer Slide Time: 04:24)



Measurement scale

Nominal scale

Data is labelled. e.g., gender, religion

Central Tendency: Mode

Hypothesis testing: Nonparametric tests such as the chi-squared test.(observed and predicted frequency)

*It is used to test if a statement regarding a population parameter is statistically significant.

Ordinal scale

Data is labelled and follow order.

e.g., Likert scale(strongly agree, agree,..)

Central Tendency: Median

Hypothesis testing: Nonparametric tests such as the Mann-Whitney U test or Wilcoxon Matched-Pairs test.

Interval scale
Ratio scale

Both scales of measurement are numerical.

Interval
Named variables
+
Ordered
+
Proportionate intervals between variables

Temperature
No absolute zero

Ratio
Named variables
+
Ordered
+
Proportionate intervals between variables
+
Can accommodate absolute zero

Height, weight etc.

Central Tendency: Mean, Median, Mode & Standard deviation

Hypothesis testing: t-test, F-test, ANOVA.

NPTEL Online Certification Courses
IIT Kharagpur

<u>Measurement scale</u>

There are different measurement scales for measuring data. In nominal scale data is labeled using categories like gender, religion, male or female and the frequency of occurrence of each category can be measured. But such data cannot be processed. The measure of central tendency used for nominal data is mode. For hypothesis testing, non-parametric tests, such as chi squared test is used. Here the observed and predicted frequency is checked to test if a statement regarding a population parameter is statistically significant or not.

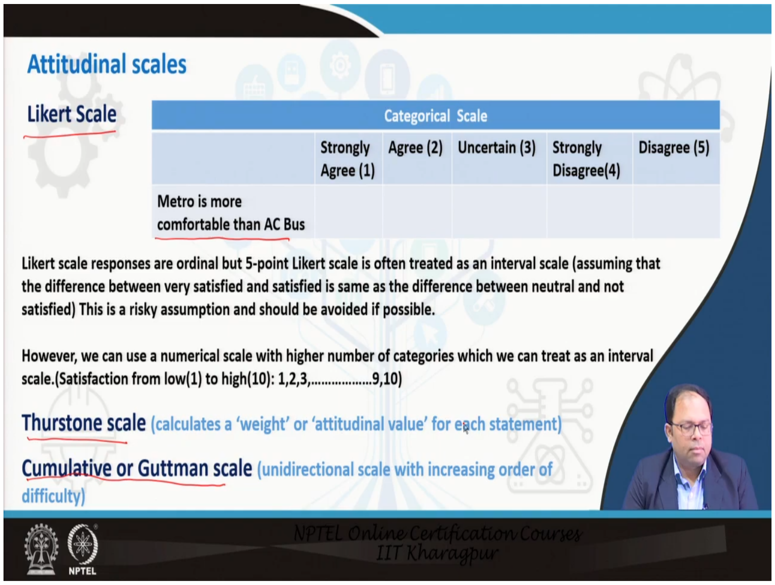
In case of ordinal scale, data is labeled following an order. For example, if there are three or four categories it can be said that the second category is higher than the first category. But such difference is qualitative and one cannot say how much higher or lower the third category is compared to the second category. One example is a Likert scale, where the scale is set like strongly agree, agree, neutral and so on. The central tendency used for measurement is median and for hypothesis testing non-parametric tests such as Mann Whitney U test or Wilcoxon Matched-Pair test is used.

The interval scale and the ratio scale are almost similar with a small difference and both the scales are used for numerical measurements. In the interval scale the variables are labeled, ordered and there is a proportionate interval between the variables. For example, on a 10 point scale the difference between 1 to 2 and 2 to 3 is equal. Also, the zero on an interval scale is relative and there is no absolute zero. Temperature is measured on an interval scale.

Whereas, in ratio scale everything else remains same but there is an absolute 0. For example, in case of height and weight there is zero weight or height.

The central tendency measure for interval and ratio scales can be mean, median or mode and the standard deviation can also be determined. For hypothesis testing t test, F test or ANOVA can be used.

(Refer Slide Time: 08:32)



Attitudinal scales

Likert Scale

Categorical Scale					
	Strongly Agree (1)	Agree (2)	Uncertain (3)	Strongly Disagree(4)	Disagree (5)
Metro is more comfortable than AC Bus					

Likert scale responses are ordinal but 5-point Likert scale is often treated as an interval scale (assuming that the difference between very satisfied and satisfied is same as the difference between neutral and not satisfied) This is a risky assumption and should be avoided if possible.

However, we can use a numerical scale with higher number of categories which we can treat as an interval scale.(Satisfaction from low(1) to high(10): 1,2,3,..................9,10)

Thurstone scale (calculates a 'weight' or 'attitudinal value' for each statement)

Cumulative or Guttman scale (unidirectional scale with increasing order of difficulty)

NPTEL Online Certification Courses
IIT Kharagpur

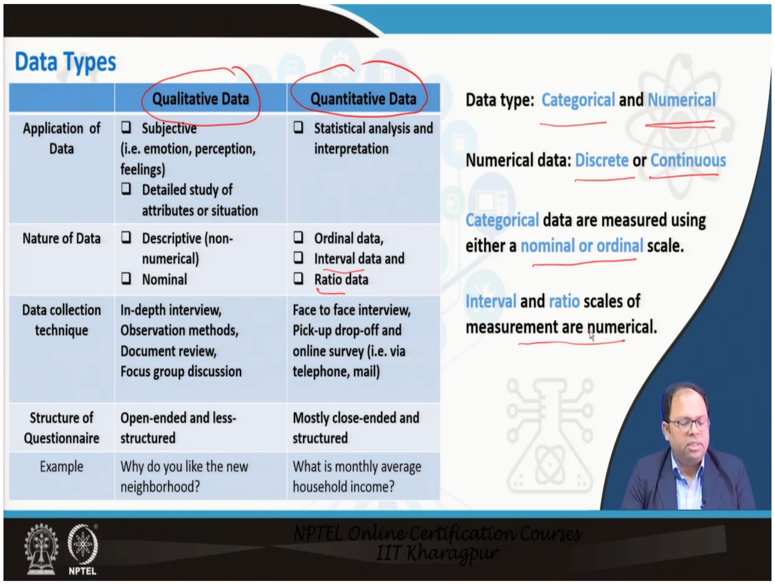
<u>Attitudinal scale</u>

This scale is used for measuring attitudes. For example one can be asked- “is metro more comfortable than air conditioned bus?” Thus, people based on their attitudes can give a certain rating. This is measured using strongly agree, agree, uncertain, strongly disagree and disagree. These attitudes can be marked 1, 2, 3, 4, 5 but the gap between 1 and 2 and 2 to 3 is not necessarily the same which means people may feel that the most difference is between strongly agree and agree, whereas from uncertain and agree the difference may be very less. But sometimes for statistical analysis the interval between two attitudes is assumed as same. This is a risky assumption and should be avoided.

Sometimes, a numerical scale with higher number of categories is used, like a 10 point scale instead of a 5 point scale. In this case, though there is difference in interval, people may assume that the difference is almost same. Thus, we can say that, a numerical scale with higher number of categories can be treated as an interval scale e.g., satisfaction from low 1 to high 10. It is still better to use a different scale instead of a Likert scale. In case a survey is conducted in India and people are asked about their satisfaction about a certain feature, then the satisfaction rating should be captured using either a 10 point scale or a 5 point scale or even a 3 point scale, to determine the appropriate scale people are comfortable with. So even though a 10 point scale may be suitable for doing statistical measurements sometimes we may find that people are not properly able to rate using a 10 point scale and may feel comfortable using just a 3 point scale or smaybe a 5 point scale. Thus, we should be very careful when using a Likert scale which is ordinal basically ordinal in nature.

In addition to the Likert scale we also have 2 more scales to measure attitude; one is called Thurstone scale and the other is called Cumulative or Guttman scale. However these are not as common as the Likert scale. In case of a Thurstone scale, every statement that is made is given a weight, though the weight may be subjective. In cumulative or Guttman scale there is a unidirectional accumulation in the interval. For example in a 10 point scale, 2 has more categories or more features than 1; 3 has got even more features than 2 but includes all features of 1 and 2 and so on. Thus, when a person chooses 8 on a 10 point scale, he/she agrees to everything till 8. So these are the 3 different attitudinal scales which are also important in transportation research.

(Refer Slide Time: 13:15)



Data Types

	Qualitative Data	Quantitative Data
Application of Data	❑ Subjective (i.e. emotion, perception, feelings) ❑ Detailed study of attributes or situation	❑ Statistical analysis and interpretation
Nature of Data	❑ Descriptive (non-numerical) ❑ Nominal	❑ Ordinal data, ❑ Interval data and ❑ Ratio data
Data collection technique	In-depth interview, Observation methods, Document review, Focus group discussion	Face to face interview, Pick-up drop-off and online survey (i.e. via telephone, mail)
Structure of Questionnaire	Open-ended and less-structured	Mostly close-ended and structured
Example	Why do you like the new neighborhood?	What is monthly average household income?

Data type: Categorical and Numerical

Numerical data: Discrete or Continuous

Categorical data are measured using either a nominal or ordinal scale.

Interval and ratio scales of measurement are numerical.

NPTEL Online Certification Courses
IIT Kharagpur

Data types

Data can also be categorized as qualitative as well as quantitative. Qualitative data can be of different kinds. For example one can measure emotion, perception, or feelings by writing about it. So it is a little bit subjective. Qualitative data can also be a detailed study of a situation. This kind of data can be described or measured using a nominal scale. It can be collected through in-depth interviews, observational methods, document reviews, focus group discussions and so on. In these cases the questions can be a bit open ended and less structured because people can have a discussion. For example, if a person is asked to describe “why do you like the new neighborhood”, he may give many reasons why he likes it. So this can be recorded and the data is essentially qualitative in nature.

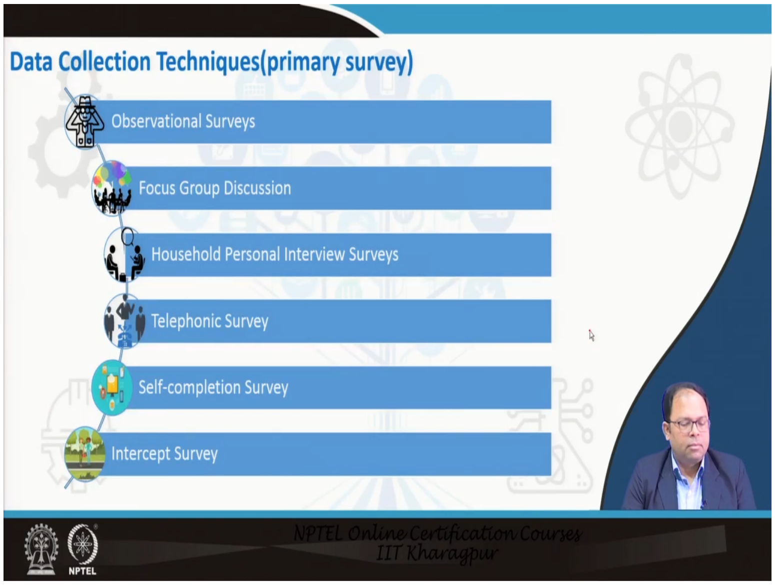
Data can also be quantitative, in which case it is open to statistical analysis and interpretations. This data can be measured using an ordinal or interval scale or it could be measured using a ratio scale. The data can be collected using face to face interview or online surveys. This kind of survey can be both close ended as well as structured. For example, a person can be specifically asked about his monthly income, on which some statistical analysis can be performed.

Data can also be classified as categorical and numerical data. In case of categorical data the entire data set can be broken into different categories or each element surveyed can be categorized. In numerical data a numerical measure can be ascribed to a particular aspect being surveyed.

Numerical data can be either discrete or continuous. Discrete means integer. For example the number of cars a family owns is an integer and therefore discrete. But the weight of a person can be 50.5 kg, i.e it can be continuous between 50 and 51.

Categorical data are measured using either nominal or ordinal scale whereas interval and ratio scale are used for measuring numerical data. So these are the different data types one can use in surveys.

(Refer Slide Time: 17:26)



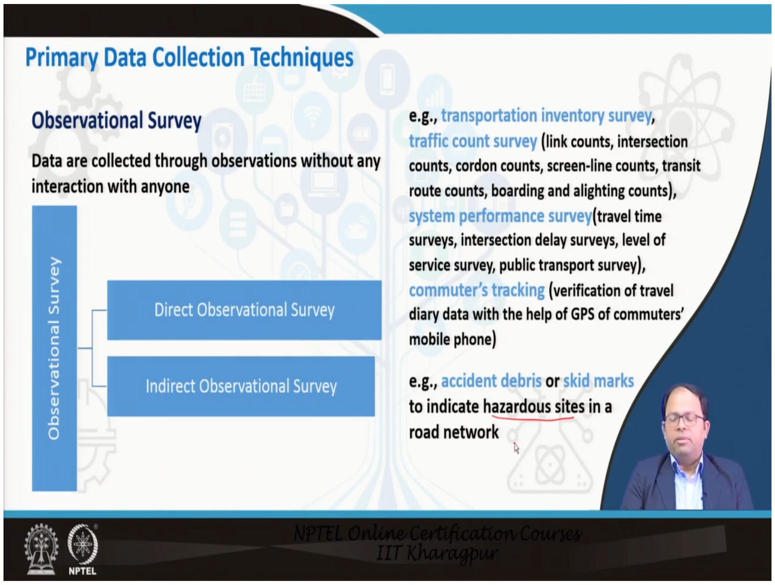
Data Collection Techniques(primary survey)

- Observational Surveys
- Focus Group Discussion
- Household Personal Interview Surveys
- Telephonic Survey
- Self-completion Survey
- Intercept Survey

Data Collection techniques

There are different data collection techniques for primary survey, such as observational surveys, focus group discussion, household personnel interviews, telephonic surveys, self completion surveys and intercept surveys.

(Refer Slide Time: 17:47)



Primary Data Collection Techniques

Observational Survey

Data are collected through observations without any interaction with anyone

Observational Survey

Direct Observational Survey

Indirect Observational Survey

e.g., transportation inventory survey, traffic count survey (link counts, intersection counts, cordon counts, screen-line counts, transit route counts, boarding and alighting counts), system performance survey(travel time surveys, intersection delay surveys, level of service survey, public transport survey), commuter's tracking (verification of travel diary data with the help of GPS of commuters' mobile phone)

e.g., accident debris or skid marks to indicate hazardous sites in a road network

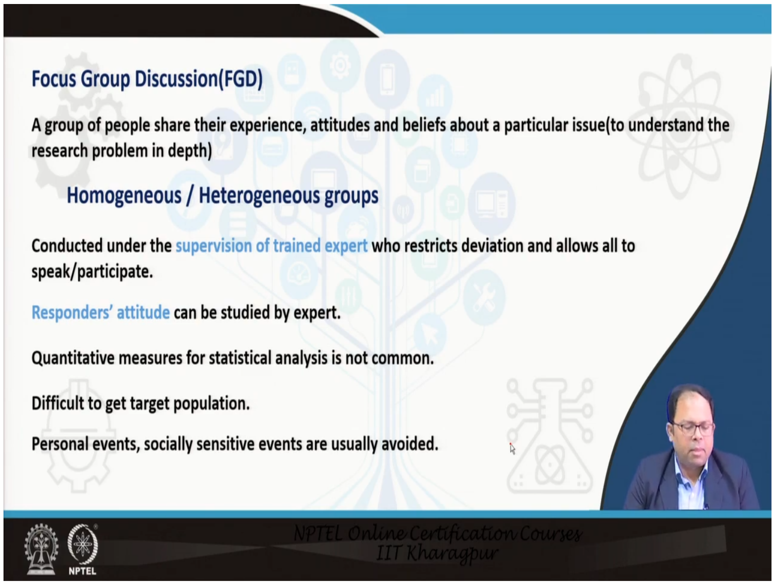
NPTEL Online Certification Courses
IIT Kharagpur

Primary data collection techniques

In observational survey, data is collected through observations and without any interactions with any particular person or individual. It can be either direct or indirect. For example, transportation inventory survey, traffic count surveys (link counts, intersection counts, cordon counts) or system performance surveys (like travel time surveys, intersection delay surveys, level of service surveys). Commuter tracking surveys like verification of the travel diary data with help of the GPS of commuter's mobile phone are all direct observational surveys.

Indirect observational survey is related to cases where instead of observing the actual phenomena the results of the phenomenon are observed. For example, accident debris or skid marks can help in marking hazardous sections of a highway and can help in identifying accident prone locations.

(Refer Slide Time: 19:17)



Focus Group Discussion(FGD)

A group of people share their experience, attitudes and beliefs about a particular issue(to understand the research problem in depth)

Homogeneous / Heterogeneous groups

Conducted under the supervision of trained expert who restricts deviation and allows all to speak/participate.

Responders' attitude can be studied by expert.

Quantitative measures for statistical analysis is not common.

Difficult to get target population.

Personal events, socially sensitive events are usually avoided.

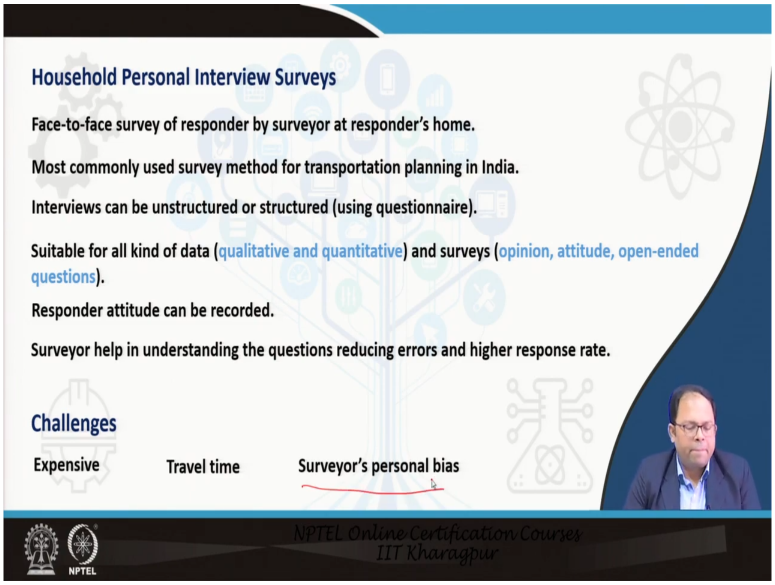
Focus group discussion

In focus group discussion a group of people share their experience, attitudes and beliefs about a particular issue. And this is done to understand the basic research problem. This also means that a group of people (homogeneous or heterogeneous) can be taken. For example, to get the feedback about an entire village, different representatives from different groups in the village can be taken. The discussion can be structured or unstructured or it can follow a script with an expert guiding the discussion.

And when people share their experiences, attitudes and beliefs then the researcher gets an in-depth understanding of the problem and using that understanding he/she can design his/her surveys or include certain variables in the surveys which would otherwise have been missed. In addition the responders' attitude can also be studied by the experts.

Quantitative measures of statistical analysis are possible but not common. People are seldom asked to rate their experience or rate their perception about certain thing during a focus group discussion. The challenge involves bringing together the target group and prevent the discussion from veering in an unrelated or socially awkward direction.

(Refer Slide Time: 21:25)



Household Personal Interview Surveys

Face-to-face survey of responder by surveyor at responder's home.

Most commonly used survey method for transportation planning in India.

Interviews can be unstructured or structured (using questionnaire).

Suitable for all kind of data (qualitative and quantitative) and surveys (opinion, attitude, open-ended questions).

Responder attitude can be recorded.

Surveyor help in understanding the questions reducing errors and higher response rate.

Challenges

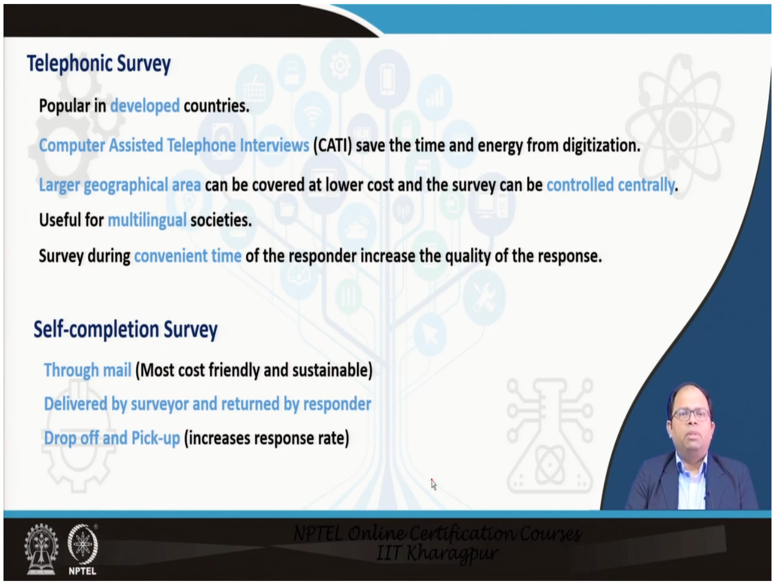
Expensive   Travel time   Surveyor's personal bias

NPTEL Online Certification Courses
IIT Kharagpur

Household personal interview surveys

This is the most common survey in transportation planning and particularly for travel diary collection. Surveyors go from door-to-door to conduct face to face interviews, which are mostly structured. Structured interviews involve the use of a questionnaire and are suitable for both qualitative and quantitative surveys. It can include either opinion survey or attitude survey or open ended questions. A questionnaire reduces ambiguity for both the surveyor and the respondent and thus reduces error. But such a survey is costly and time consuming. There is also a chance of the surveyor's bias creeping in.

(Refer Slide Time: 23:07)



Telephonic Survey

Popular in developed countries.

Computer Assisted Telephone Interviews (CATI) save the time and energy from digitization.

Larger geographical area can be covered at lower cost and the survey can be controlled centrally.

Useful for multilingual societies.

Survey during convenient time of the responder increase the quality of the response.

Self-completion Survey

Through mail (Most cost friendly and sustainable)

Delivered by surveyor and returned by responder

Drop off and Pick-up (increases response rate)

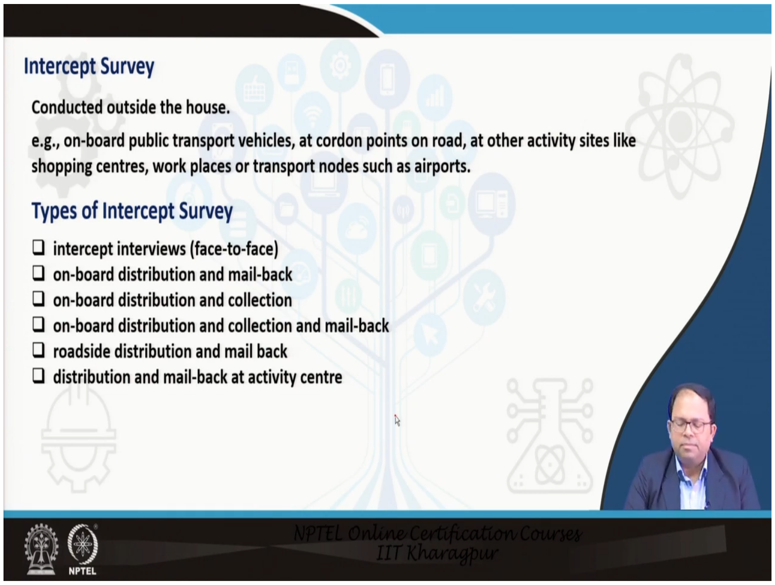
NPTEL Online Certification Courses
IIT Kharagpur

Telephonic surveys

Telephonic survey is an alternative to the household survey and is very popular in the developed world. These surveys can be computer assisted with responses going directly into a database and can thus save time and energy from digitization. This kind of survey can be centrally controlled for a very large area with a multi-lingual society and thus saves time and cost. Responses can be sought at convenient times improving response rate and overall quality of response.

Self completion surveys are the ones where a person has to fill up the responses himself. And this could be done in 3 ways. It could be either done through mail which is cost effective. Or it could be delivered by the surveyor to each household and the responder has to return it on their own. Or the surveyor drops-off and picks up the response directly from the household, in which case the probability of the response increases.

(Refer Slide Time: 24:58)



Intercept Survey

Conducted outside the house.

e.g., on-board public transport vehicles, at cordon points on road, at other activity sites like shopping centres, work places or transport nodes such as airports.

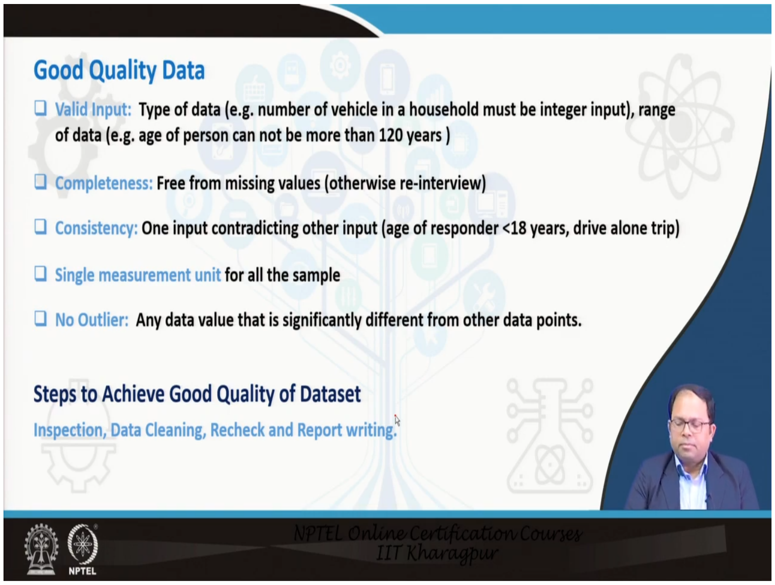
Types of Intercept Survey

- ❑ intercept interviews (face-to-face)
- ❑ on-board distribution and mail-back
- ❑ on-board distribution and collection
- ❑ on-board distribution and collection and mail-back
- ❑ roadside distribution and mail back
- ❑ distribution and mail-back at activity centre

Intercept survey

Sometimes it is difficult to conduct surveys inside a house and to get hold of a target responder. For example, in a survey of public transit users it is better to conduct the survey on a public transit vehicle. Similarly for shopping centers, workplaces or transport nodes such as airports it is better to conduct the surveys at the location. These are called intercept surveys. There are different kinds of intercept surveys- face to face, on-board distribution of the forms and mail back of those forms, on-board distribution and collection after an interval, on-board distribution and collection or mail-back, and roadside distribution and mail back at activity center.

(Refer Slide Time: 26:27)



Good Quality Data

- □ Valid Input: Type of data (e.g. number of vehicle in a household must be integer input), range of data (e.g. age of person can not be more than 120 years)
- □ Completeness: Free from missing values (otherwise re-interview)
- □ Consistency: One input contradicting other input (age of responder <18 years, drive alone trip)
- □ Single measurement unit for all the sample
- □ No Outlier: Any data value that is significantly different from other data points.

Steps to Achieve Good Quality of Dataset

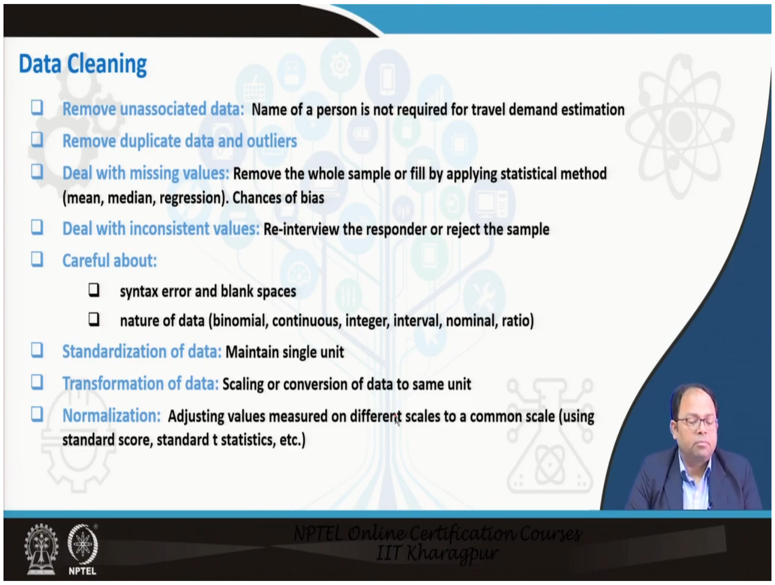
Inspection, Data Cleaning, Recheck and Report writing.

NPTEL Online Certification Courses
IIT Kharagpur

Quality of data

Surveys should ensure good quality of data. So type of data is important. And also the range of data is important. For example, if suddenly a person records that age is 120 years then that means it is a wrong input. So these are the things that should be checked. An incomplete survey cannot be accepted and has to be rejected or repeated by the surveyor. Also data inconsistency has to be rectified. For example, if an under eighteen respondent says he drives his own car to school, it is probably not true. The units of quantitative data, if any, have to be uniform too. Outliers in data (i.e. data points vastly different from the rest) have to be removed too. So, the steps to achieve good quality data is inspection, cleaning and recording the changes made to the data while cleaning. The last step ensures that any problem with the data in the future can be traced back.

(Refer Slide Time: 28:51)



Data Cleaning

- ❑ Remove unassociated data: Name of a person is not required for travel demand estimation
- ❑ Remove duplicate data and outliers
- ❑ Deal with missing values: Remove the whole sample or fill by applying statistical method (mean, median, regression). Chances of bias
- ❑ Deal with inconsistent values: Re-interview the responder or reject the sample
- ❑ Careful about:
  - ❑ syntax error and blank spaces
  - ❑ nature of data (binomial, continuous, integer, interval, nominal, ratio)
- ❑ Standardization of data: Maintain single unit
- ❑ Transformation of data: Scaling or conversion of data to same unit
- ❑ Normalization: Adjusting values measured on different scales to a common scale (using standard score, standard t statistics, etc.)

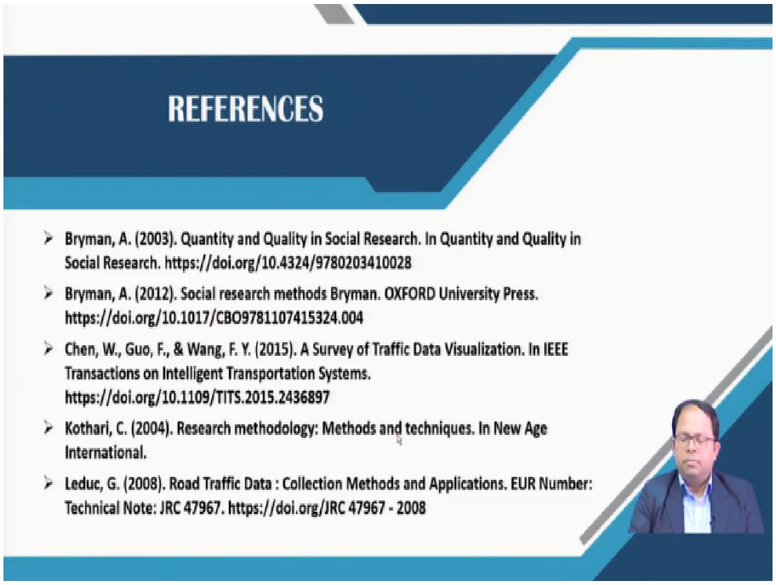
Data Cleaning

Data cleaning is a very important step in the survey process that has to be completed before analysis is done. It involves removing inconsistent and irrelevant data, like name of a person in a travel diary survey. Post curve fitting using a regression analysis data points lying faraway are removed. These are called outliers. Quartile measure can too be used for removal of outliers.

Data gaps may be filled using mean or median of the dataset. But such an exercise can lead to a bias with too many data points close to the central tendency. Then there is the need for standardization of the data collected, like reconciliation of various units of measurement followed by scaling or normalization. Data normalizing means adjusting values measured on different scales to a common scale and this can be done using standard scores, standard t statistics and so on. Finally, the kind of analysis to be carried out on the data is based on the nature of the data, which can be binomial, continuous, integer, interval, nominal, etc.

(Refer Slide Time: 32:27)

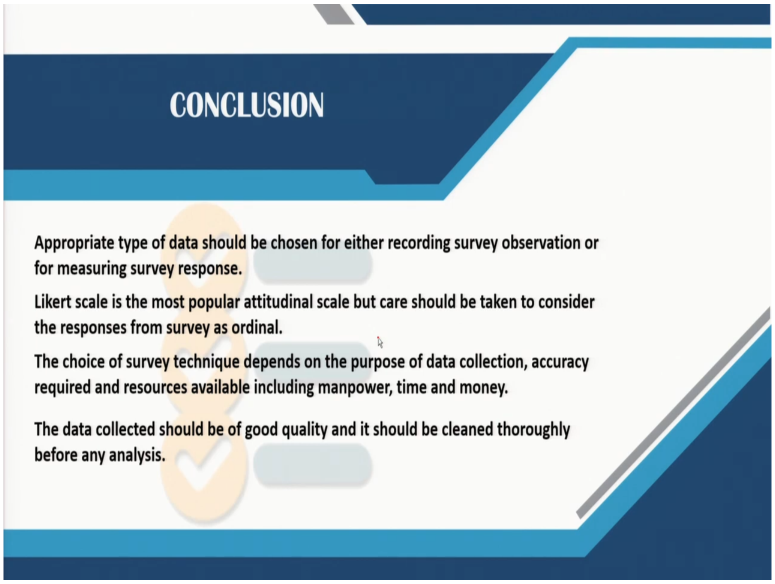
References



REFERENCES

- 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
- Chen, W., Guo, F., & Wang, F. Y. (2015). A Survey of Traffic Data Visualization. In IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2015.2436897
- Kothari, C. (2004). Research methodology: Methods and techniques. In New Age International.
- Leduc, G. (2008). Road Traffic Data : Collection Methods and Applications. EUR Number: Technical Note: JRC 47967. https://doi.org/JRC 47967 - 2008

(Refer Slide Time: 32:35)



CONCLUSION

Appropriate type of data should be chosen for either recording survey observation or for measuring survey response.

Likert scale is the most popular attitudinal scale but care should be taken to consider the responses from survey as ordinal.

The choice of survey technique depends on the purpose of data collection, accuracy required and resources available including manpower, time and money.

The data collected should be of good quality and it should be cleaned thoroughly before any analysis.

So in conclusion it can be said that appropriate type of data should be chosen for either recording survey observation or for measuring survey response. Likert scale is the most popular attitudinal scale but care should be taken to consider the responses from the survey as ordinal. The choice of survey technique depends on the purpose of data collection, accuracy required and resources available including manpower, time and money. And the data collected should be of good quality and it should be cleaned thoroughly before any analysis. Thank you.