COLLECTING AND SUMMARIZING DATA

Module 2

Collecting Data
Lesson 1

Using Numbers to Describe Data
Lesson 2

OVERVIEW
OVERVIEW
OVERVIEW

Data collection brings together raw data, but this has to be processed to give useful information. The amount of detail in data sometimes obscures underlying patterns and data reduction clears away the detail and highlights important features and patterns. It enables a view of the data that is concise, but still accurate. There are two approaches to summarizing and presenting data.

OBJECTIVES

- To appreciate the importance of data collection
- To discuss the amount of data to be collected
- To discuss the aims of data reduction and presentation
- To design tables of numerical data
- To draw frequency distributions of data
- To use graphs to show the relationship between two variables
- To draw charts
- To appreciate the need for numerical measures of data
- To understand measures of location
- To understand the arithmetic mean, median and mode of data
- To understand the purpose of index numbers

Collecting Data

x

Framework is a formal structure of giving you a list of the key elements that key factors we use in enterprise architecture.

Data and Information

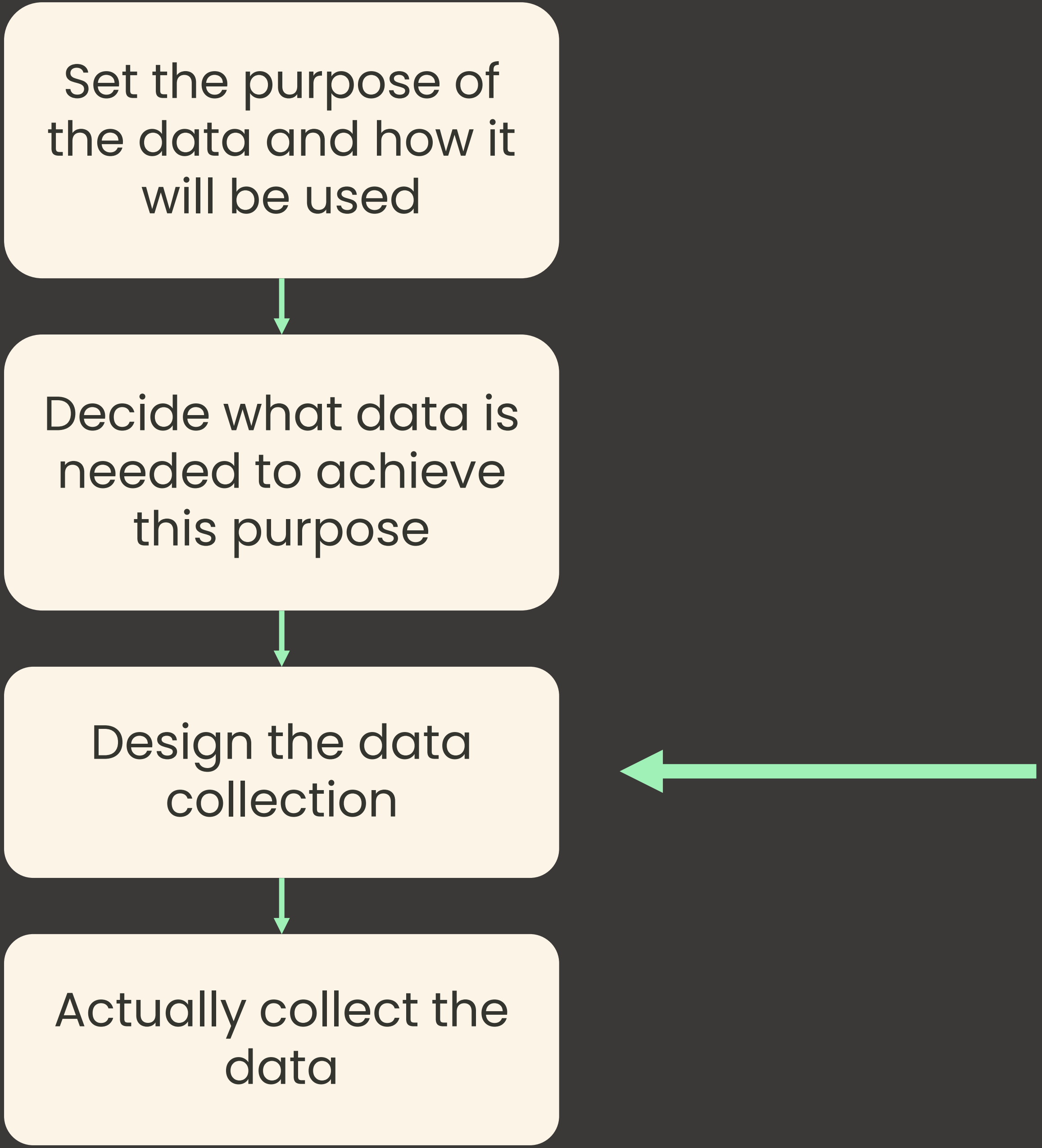
Managers need relevant information for their decisions. To get this, they start with data collection, then process the data to give information, and present the results in the best formats. Data collection is essential in every organization, because it starts the process of decision-making – and without proper data collection, managers cannot make informed decisions

Data collection does not happen by chance, but needs careful planning.

1. Define the purpose of the data.
2. Describe the data you need to achieve this purpose.
3. Check available secondary data to see how useful it is.
4. Define the population and sampling frame to give primary data.
5. Choose the best sampling method and sample size.
6. Identify an appropriate sample.
7. Design a questionnaire or other method of data collection.
8. Run a pilot study and check for problems.
9. Train interviewers, observers or experimenters.
10. Do the main data collection.
11. Do follow-up, such as contacting non-respondents.
12. Analyze and present the results.

Planning Data Collection

x



Set the purpose of the data and how it will be used

Decide what data is needed to achieve this purpose

Design the data collection

Actually collect the data

Includes:

- the amount of data
- sources of data
- means of collection

Amount of Data

Three important questions for data collection are the amount of data, its source, and the means of collection.

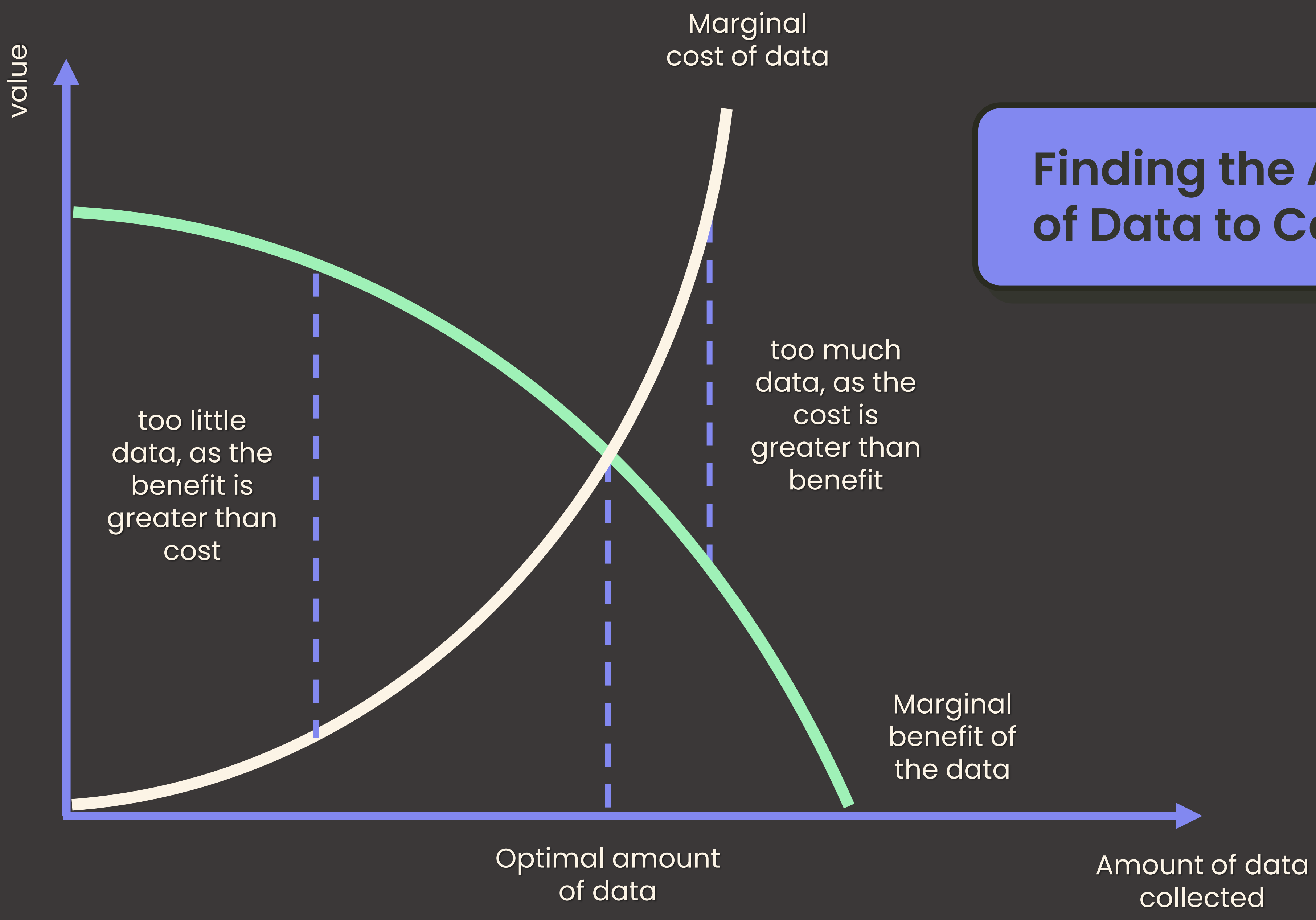
A marginal cost as the extra cost of collecting one more bit of data, and this rises with the amount of data collected.

The marginal benefit of data – which is the benefit from the last bit collected – falls with the amount collected.

In principle, you collect the amount where the marginal cost equals the marginal benefit.
If you collect less than this, you lose potential benefit because the cost of collection is less than the benefit; if you collect more data than this, you waste resources because the cost of collection is more than the benefit.

Finding the Amount of Data to Collect

x



Types of Data

Nominal Data

The kind that we really cannot quantify with any meaningful units.

Ordinal Data

One step more quantitative than nominal data. Here we can rank the categories of observations into some meaningful order.

Cardinal Data

Has some attribute that can be measured directly.

Primary and Secondary Data

Primary Data

This is often characterized as **field research** when you actually go out and collect data yourself. It is a new data collected by an organization itself for a specific purpose.

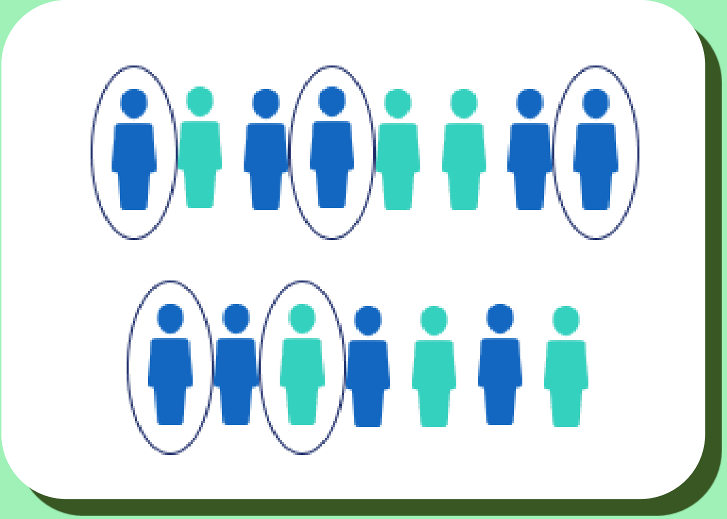
Secondary Data

Characterized as **desk research** when you look for data that someone else has already collected. It is an existing data that was collected by other organizations or for other purposes.

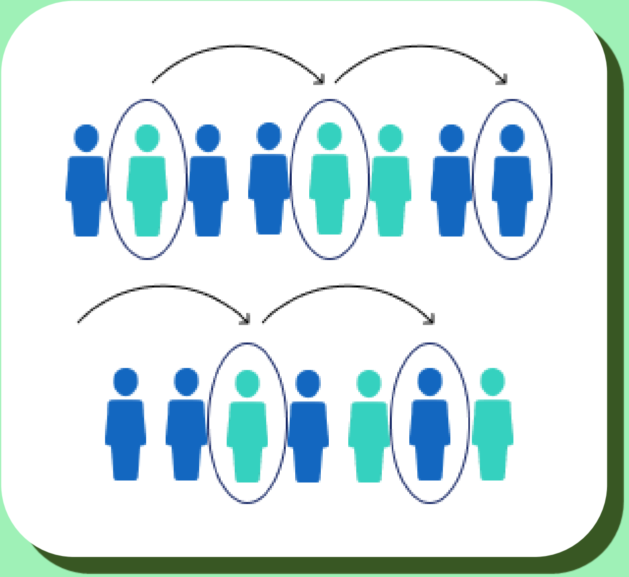
Using Samples to Collect Data

A population consists of all the people or items that could supply data. These are listed in a sampling frame. It can be difficult to choose the right population and find a sampling frame.
It is usually too expensive, time-consuming and difficult to collect data from the whole population – giving a census. The alternative collects data from a representative sample of the population and uses this to estimate values for the whole population.

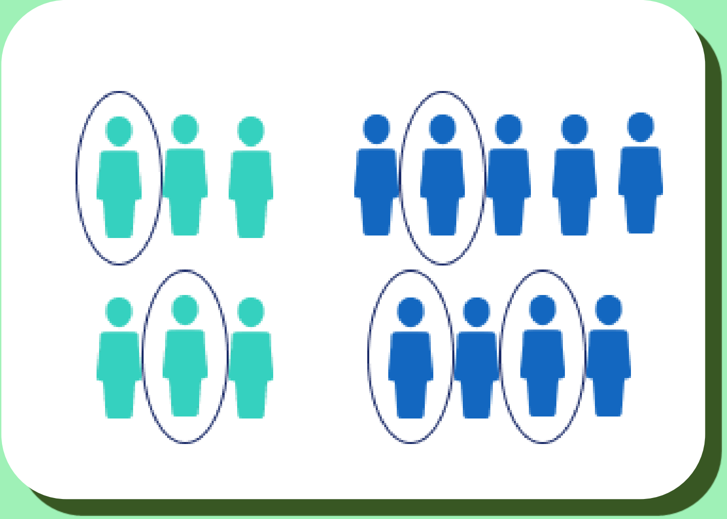
SIMPLE
RANDOM
SAMPLE



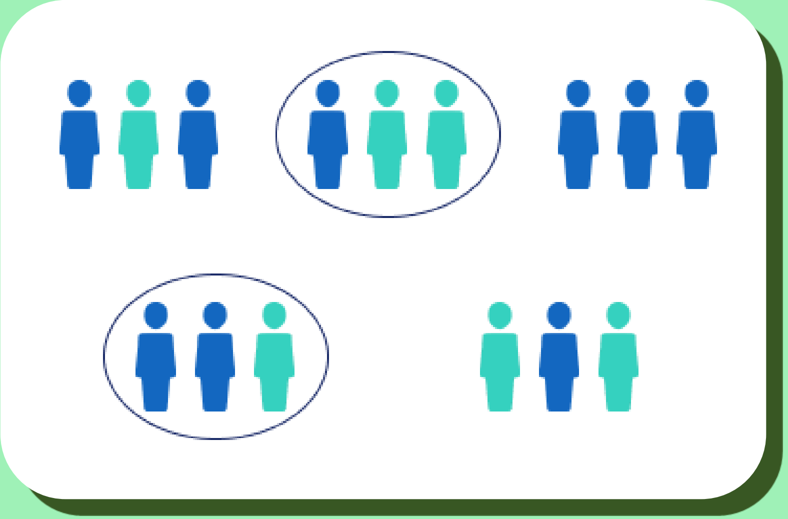
SYSTEMATIC
SAMPLE



STRATIFIE
D
SAMPLE

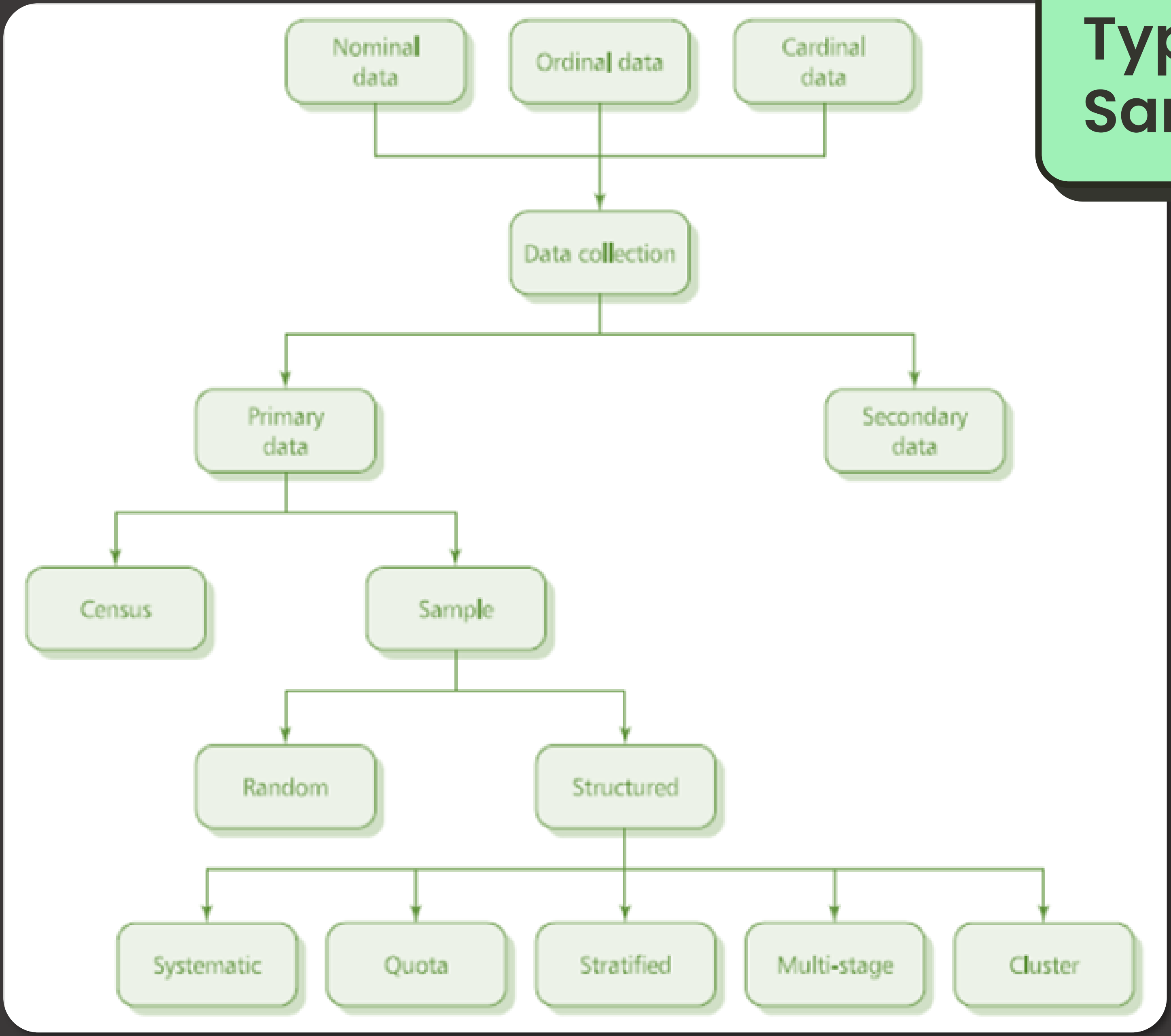


CLUSTER
SAMPLE

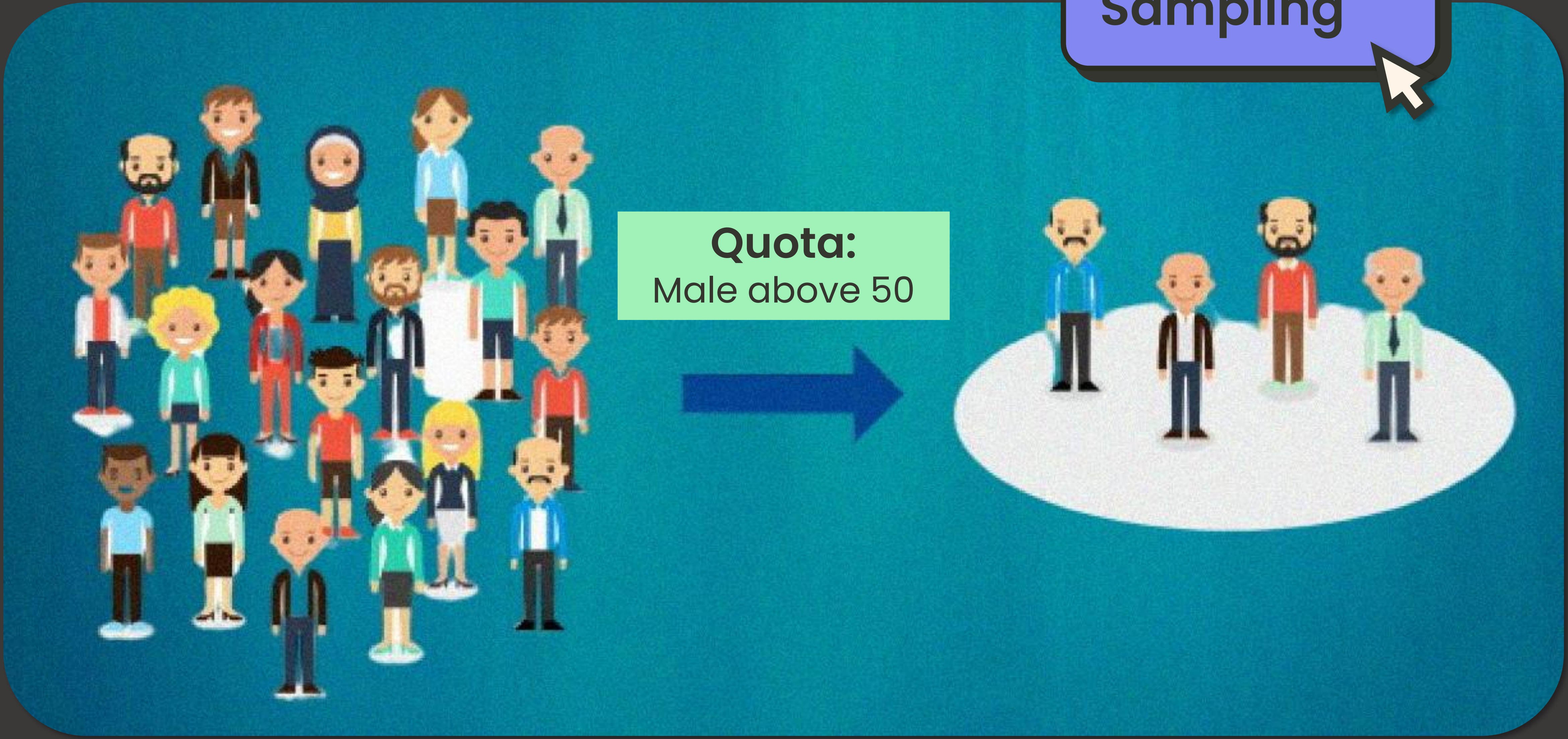


Types of Samples

x



Quota Sampling



Quota:
Male above 50

Organizing Data Collection

Observation

A way of gathering data by watching behavior, events, or noting physical characteristics in their natural setting. It can be overt or covert. It can also be either direct or indirect.

Questionnaires

Can be administered through personal interview, telephone interview, the Internet, postal survey, panel survey or longitudinal survey.

How to Plan for Observations

- **Determine the focus**. Think about the evaluation question(s) you want to answer through observation and select a few areas of focus for your data collection.
- **Design a system for data collection**. Think about the specific items for which you want to collect data and then determine how you will collect the information you need. There are three primary ways of collecting observation data: recording sheets and checklist; observation guides, and field notes.
- **Select the sites**. Select an adequate number of sites to help ensure they are representative of the larger population and will provide an understanding of the situation you are observing

How to Plan for Observations

- **Select the observers**. You may choose to be the only observer or you may want to include others in conducting observations.
- **Train the observers**. It is critical that the observers are well trained in your data collection process to ensure high quality and consistent data.
- **Time your observations appropriately**. Programs and processes typically follow a sequence of events. It is critical that you schedule your observations so you are observing the components of the activity that will answer your evaluation questions. This requires advance planning.

Guidelines for Designing Questionnaires

- A questionnaire should ask a series of related questions and should follow a logical sequence.
- Make the questionnaire as short as possible.
- Questions should be short, simple, unambiguous, and easy to understand and phrased in everyday terms.
- Even simple changes to phrasing can give very different results.
- Avoid leading questions. Such questions encourage conformity rather than truthful answers.
- Use phrases that are as neutral as possible.
- Phrase all personal questions carefully.

Guidelines for Designing Questionnaires

- Do not give warnings – a question that starts ‘We understand if you do not want to answer this, but . . .’ will discourage everyone from answering.
- Avoid vague questions.
- Ask positive questions rather than a less definite one.
- Avoid hypothetical questions. Any answer will be speculative and probably not based on any real thought.
- Avoid asking two or more questions in one.

Guidelines for Designing Questionnaires

- Open questions collect general views, but they favor the articulate and quick thinking, and are difficult to analyze.
- Ask questions with preceded answers, with respondents choosing the most appropriate answer from a set of alternatives. There are many formats for these.
- Be prepared for unexpected effects, such as sensitivity to the color and format of the questionnaire, or different types of interviewer getting different responses.
- Always run a pilot survey before starting the whole survey. This is the only way to identify problems and improve the questionnaire design

Guidelines for Designing Questionnaires

x

ERRORS

- failure to identify an appropriate population
- choosing a sample that does not represent this population
- mistakes in contacting members of the sample
- mistakes in collecting data from the sample
- introducing bias from non-respondents
- mistakes made during data analysis
- drawing invalid conclusions from the analysis

Using Numbers To Describe Data

x

MEASURING DATA

Numerical measures of data give more objective and accurate descriptions.

Measure of Location

to show where the center of the data is, giving some kind of typical or average value

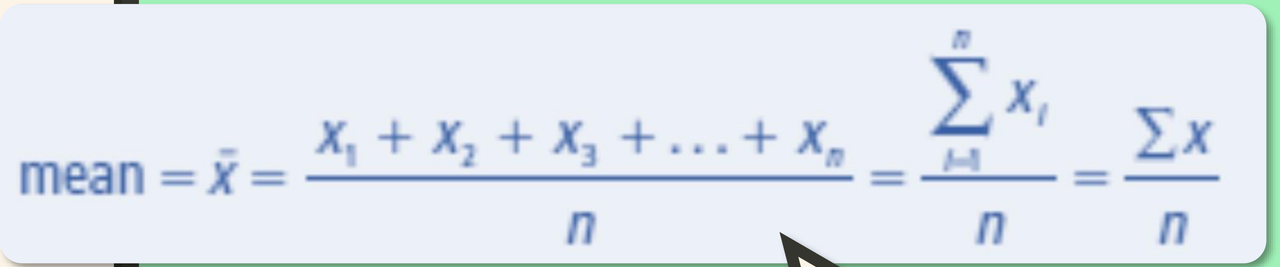
Measure of Spread

to show how the data is scattered around this center, giving an idea of the range of values.

MEASURE OF LOCATION

Arithmetic Mean

- the most widely used measure, giving an average value
- add all the values together to get the sum
- divide this sum by the number of values to get the mean.



$$\text{mean} = \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

MEASURE OF LOCATION

Median

x

- arranging the values in order of size
- counting the number of values
- identifying the middle value – which is the median.

Mode

x

- drawing a frequency distribution of the data
- identifying the most frequent value – which is the mode.

Choice of Measure

x

- the mean is the simple average
- the median is the middle value
- the mode is the most frequent value.

Advantages and Disadvantages

MEDIAN

x

Advantage

- Simple to understand and easy to calculate
- The middle part of the set, not affected by the extreme values.

Disadvantage

- Need to be arranged, if the set contains a large amount of quantity, the process is slow.
- Cannot be rigidly defined, if you happen to have even number of items, median cannot be exactly found.

Advantages and Disadvantages

MODE

Advantage

- Simple to understand and easy to calculate
- Not affected by extremely large or small values.
- Can be located just by inspection
- Useful for qualitative data

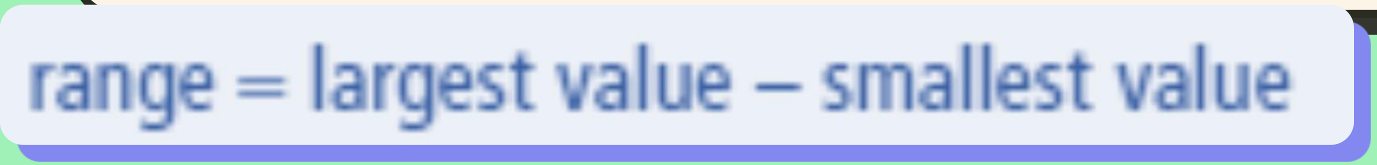
Disadvantage

- Not well defined
- Not based on all the values
- Not capable of further mathematical treatment
- Sometimes the data has one or more than one mode and sometimes the data has no mode at all.

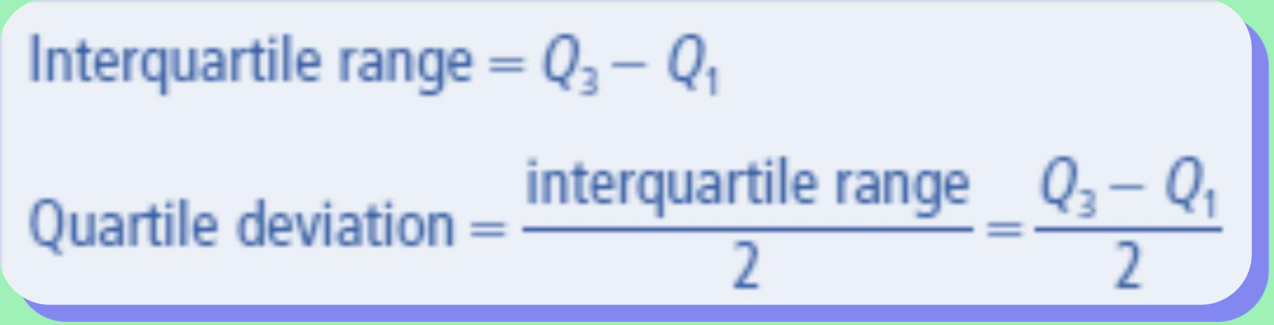
MEASURE OF SPREAD

Range and Quartiles

- The simplest measure of spread is the range, which is the difference between the largest and smallest values in a set of data.
- The broader the range, the more spread out the data.



$$\text{range} = \text{largest value} - \text{smallest value}$$



$$\text{Interquartile range} = Q_3 - Q_1$$

$$\text{Quartile deviation} = \frac{\text{interquartile range}}{2} = \frac{Q_3 - Q_1}{2}$$

MEASURE OF SPREAD

Mean Absolute Deviation

x

- The deviation is the difference between a value and the mean.
- A basic measure gives the mean absolute deviation. Alternatively, we can square the deviations and calculate the mean squared deviation – or the variance.

$$\text{mean deviation} = \frac{\sum(x - \bar{x})}{n}$$

$$\text{mean absolute deviation} = \frac{\sum \text{ABS}(x - \bar{x})}{n}$$

$$\text{MAD} = \frac{\sum|x - \bar{x}|}{n}$$

MEASURE OF SPREAD

Variance and Standard Deviation x

The square root of the variance is the standard deviation, which is the most widely used measure of spread. We can usually estimate the number of observations expected within a certain number of standard deviations of the mean.

$$\text{variance} = \frac{\sum (x - \bar{x})^2}{n}$$

The standard deviation is used for other analyses, such as the **coefficient of variation** – gives a relative view of spread, and the **coefficient of skewness** – describes the shape of a distribution.

$$\text{standard deviation} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$= \sqrt{\text{variance}}$$