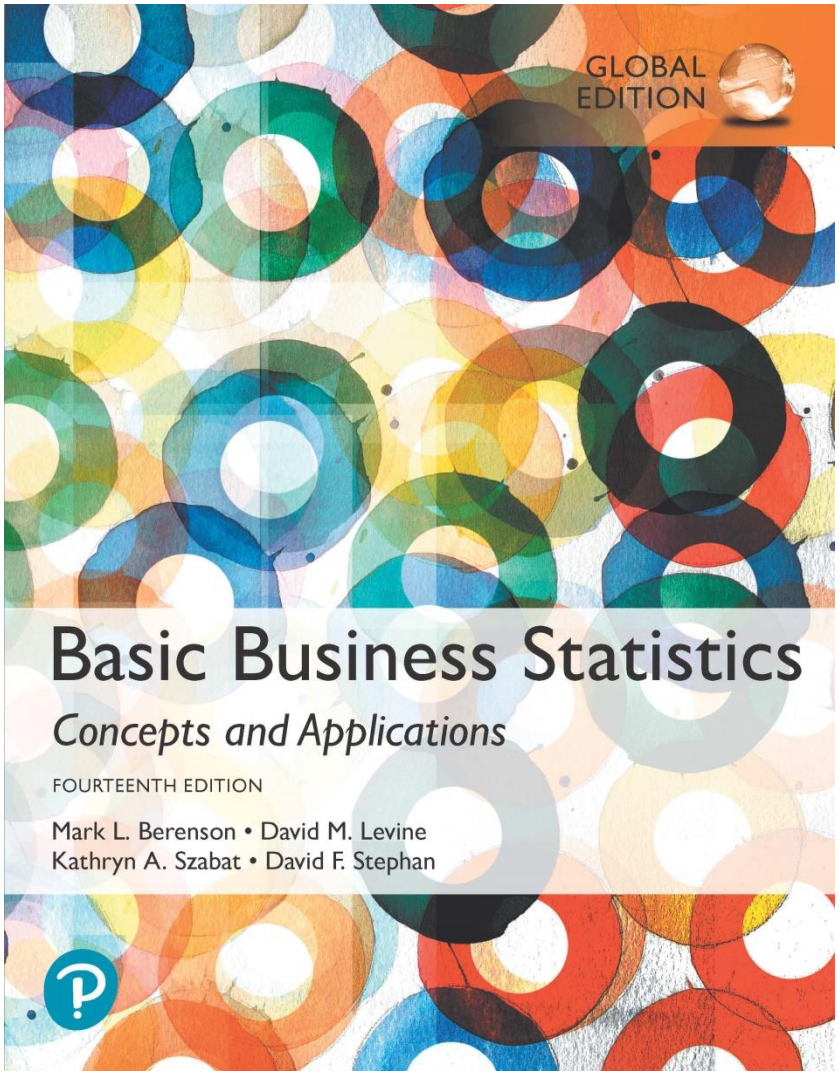


Chapter 1

Defining and Collecting Data

Objectives

In this chapter you learn:

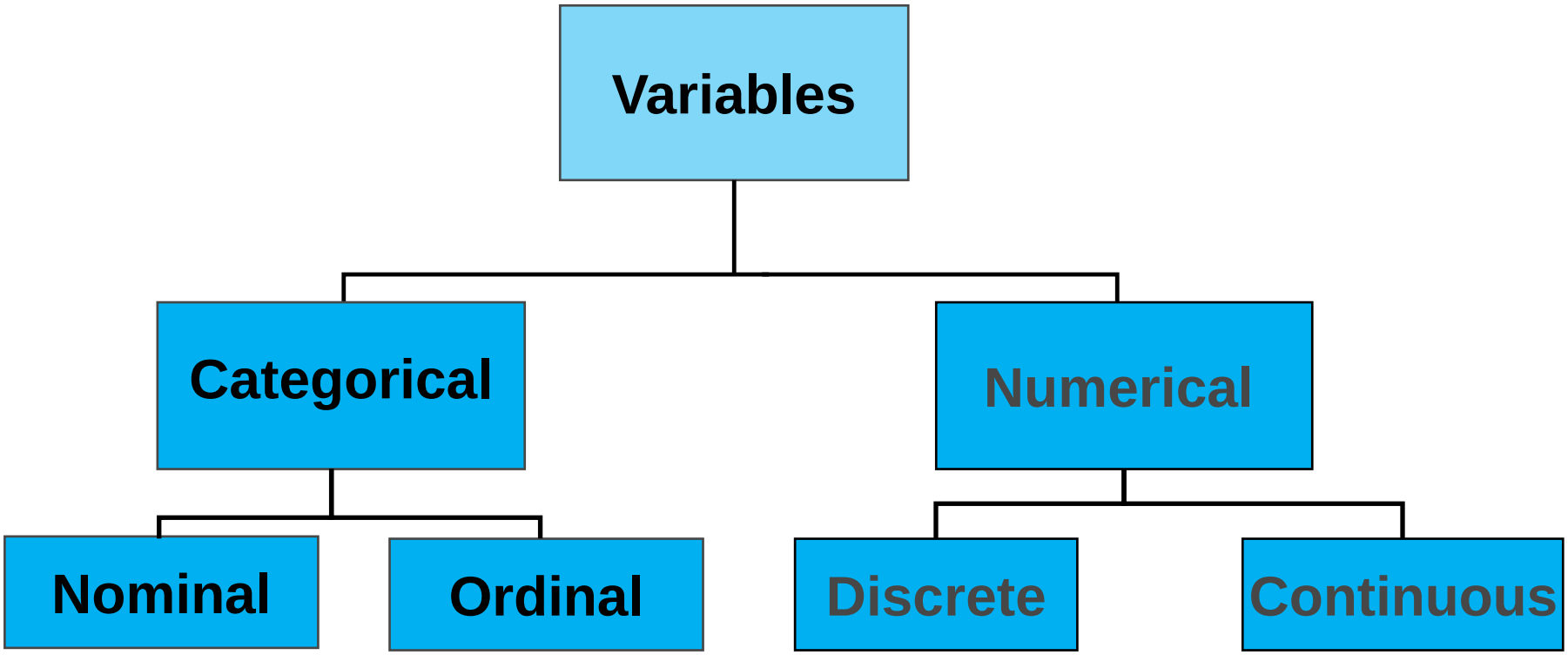
- To understand issues that arise when defining variables.
- How to define variables.
- To understand the different measurement scales.
- How to collect data.
- To identify different ways to collect a sample.

Classifying Variables By Type

- **Categorical** (***qualitative***) variables take categories as their values such as “yes”, “no”, or “blue”, “brown”, “green”.

- **Numerical** (***quantitative***) variables have values that represent a counted or measured quantity.
  - **Discrete** variables arise from a ***counting process.***
  - **Continuous** variables arise from a ***measuring process.***

Types of Variables



Examples:

- Marital Status
- Political Party
- Eye Color

(Defined Categories)

Examples: Ratings

- Good, Better, Best
- Low, Med, High

(Ordered Categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Data Is Collected From Either A Population or A Sample

POPULATION

A **population** contains all of the items or individuals of interest that you seek to study.

SAMPLE

A **sample** contains only a portion of a population of interest.

Population vs. Sample

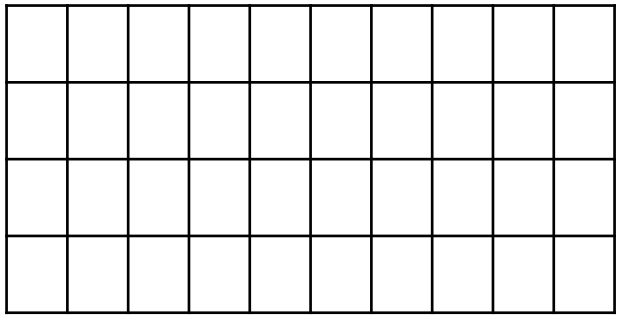
Population

All the items or individuals about which you want to draw conclusion(s).

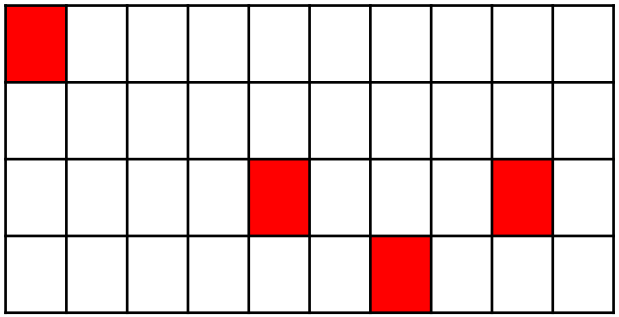
Sample

A portion of the population of items or individuals.

A Population of Size 40



A Sample of Size 4



Collecting Data Via Sampling Is Used When Doing So Is

- Less time consuming than selecting every item in the population.
- Less costly than selecting every item in the population.
- Less cumbersome and more practical than analyzing the entire population.

Parameter or Statistic?

- A **population parameter** summarizes the value of a specific variable for a population.

- A **sample statistic** summarizes the value of a specific variable for sample data.

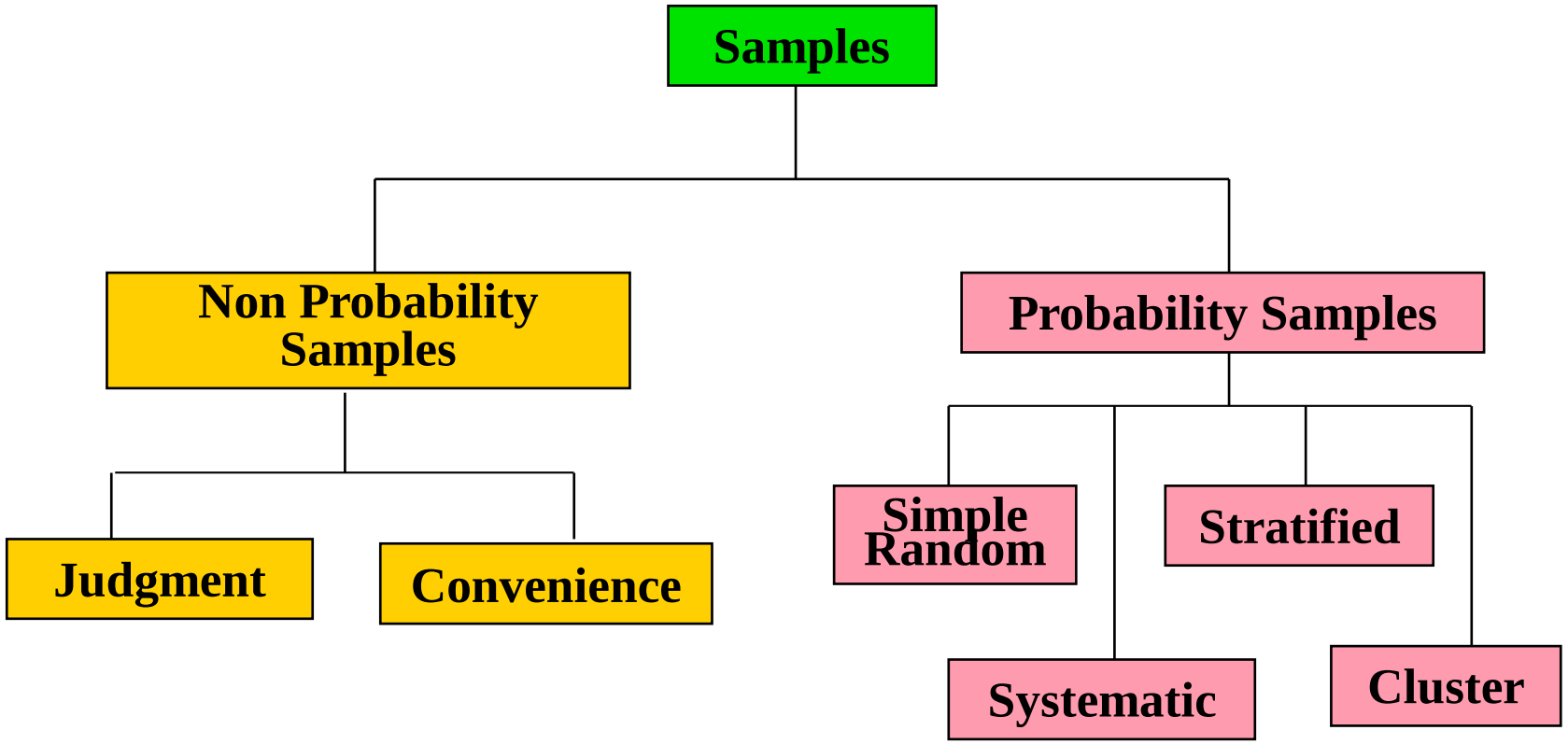
Sources of Data

- **Primary Sources**: The data collector is the one using the data for analysis:
  - Data from a political survey.
  - Data collected from an experiment.
  - Observed data.
- **Secondary Sources**: The person performing data analysis is not the data collector:
  - Analyzing census data.
  - Examining data from print journals or data published on the internet.

A Sampling Process Begins With A Sampling Frame

- The sampling frame is a listing of items that make up the population.
- Frames are data sources such as population lists, directories, or maps.
- Inaccurate or biased results can result if a frame excludes certain groups or portions of the population.
- Using different frames to generate data can lead to dissimilar conclusions.

Types of Samples

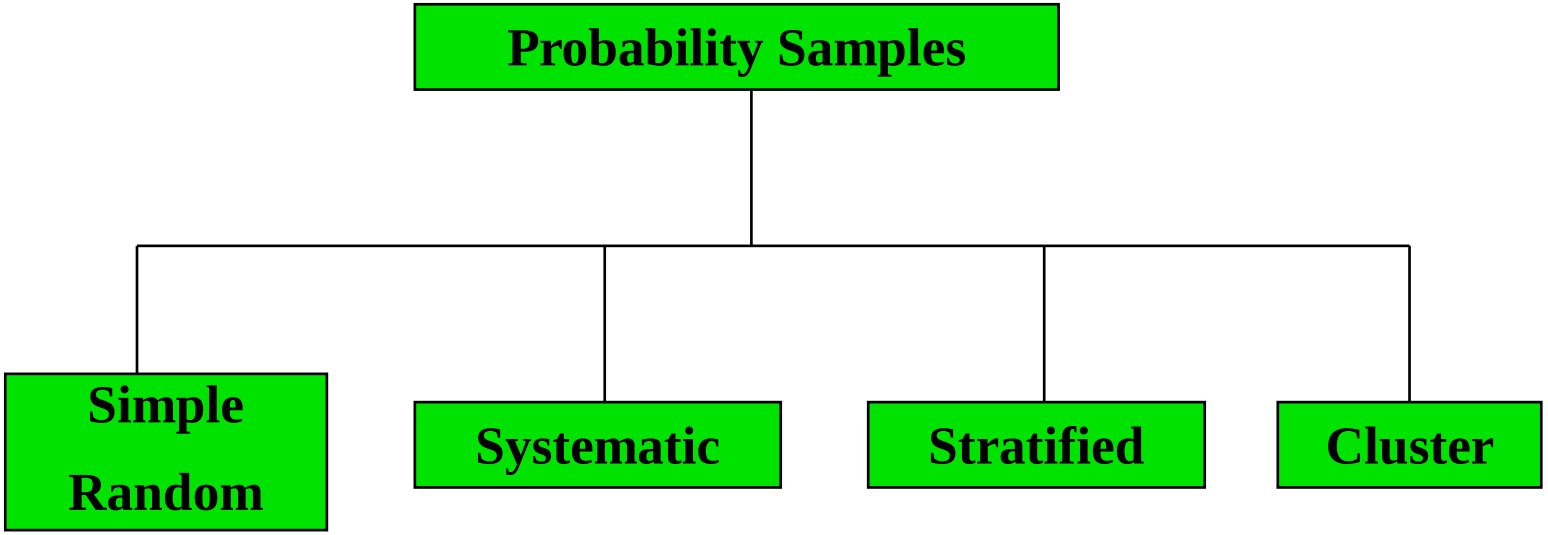


Types of Samples: Nonprobability Sample

- In a nonprobability sample, items included are chosen without regard to their probability of occurrence.
  - In **convenience sampling**, items are selected based only on the fact that they are easy, inexpensive, or convenient to sample.
  - In a **judgment sample,** you get the opinions of pre-selected experts in the subject matter.

Types of Samples: Probability Sample

- In a **probability sample**, items in the sample are chosen on the basis of known probabilities.



Probability Sample: Simple Random Sample

- **Every individual** or item from the frame has an **equal chance** of being selected.
- Selection may be with replacement (selected individual is returned to frame for possible reselection) or without replacement (selected individual isn't returned to the frame).
- Samples obtained from table of random numbers or computer random number generators.

Selecting a Simple Random Sample Using A Random Number Table

Sampling Frame For Population With 850 Items

Item Name	Item #
Bev R.	001
Ulan X.	002
.	.
.	.
.	.
.	.
Joann P.	849
Paul F.	850

Portion Of A Random Number Table

49280 88924 35779 00283 81163 07275
11100 02340 12860 74697 96644 89439
09893 23997 20048 49420 88872 08401

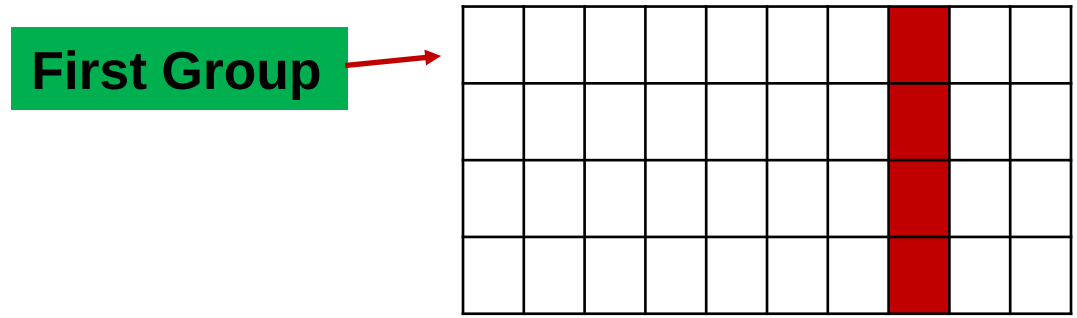
The First 5 Items in a simple random sample

Item # 492
Item # 808
Item # 892 -- does not exist so ignore
Item # 435
Item # 779
Item # 002

Probability Sample: Systematic Sample

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group
- Select every kth individual thereafter

N = 40
n = 4
k = 10

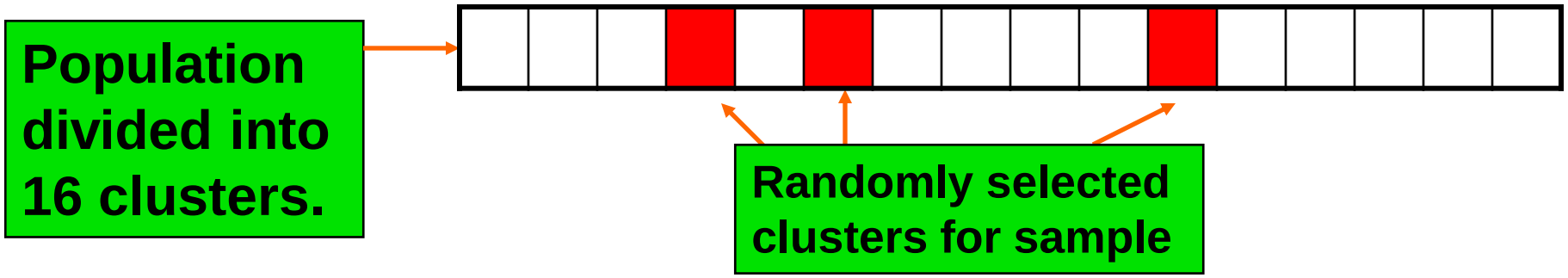


Probability Sample: Stratified Sample

- Divide population into two or more subgroups (called *strata*) according to some common characteristic.
- A **simple random sample is selected from each subgroup**, with **sample sizes proportional to strata sizes**.
- Samples from subgroups are combined into one.
- This is a common technique when sampling population of voters, stratifying across racial or socio-economic lines.

Probability Sample
Cluster Sample

- Population is divided into several “clusters,” each representative of the population.
- A simple random sample of clusters is selected.
- All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique.
- A common application of cluster sampling involves election exit polls, where certain election districts are selected and sampled.



Probability Sample: Comparing Sampling Methods

- Simple random sample and Systematic sample:
  - Simple to use.
  - May not be a good representation of the population's underlying characteristics.
- Stratified sample:
  - Ensures representation of individuals across the entire population.
- Cluster sample:
  - More cost effective.
  - Less efficient (need larger sample to acquire the same level of precision).

Data Cleaning Is An Important Data Preprocessing Task Prior To Analysis

Data cleaning corrects irregularities in the data:

- Invalid variable values, including:
  - Non-numerical data for numerical variable.
  - Invalid categorical values for a categorical variable.
  - Numeric values outside a defined range.
- Coding errors, including:
  - Inconsistent categorical values.
  - Inconsistent case for categorical values.
  - Extraneous characters.
- Data integration errors, including:
  - Redundant columns.
  - Duplicated rows.
  - Differing column lengths.
  - Different units of measure or scale for numerical variables.

Chapter Summary

In this chapter we have discussed:

- Understanding issues that arise when defining variables.
- How to define variables.
- Understanding the different measurement scales.
- How to collect data.
- Identifying different ways to collect a sample.
- Understanding the issues involved in data preparation.
- Understanding the types of survey errors.