Chapter 2

Collection and Presentation of Data

Outline of Chapter 2

- Preliminaries
- Methods of Data Collection
  - Common Methods of Data Collection
  - The Questionnaire
- **Sampling**
  - **Methods of Probability Sampling**
  - **Methods of Nonprobability Sampling**
- Tabular and Graphical Presentation
- The Frequency Distribution
- The Stem-and-Leaf Display

Reference: Chapters 2-4,12 of Elementary Statistics by ACS

Part 2

Sampling and Sampling Techniques

Recall

What is population?

What is a sample?

Motivation

Basic Concepts

- In **complete enumeration** or a **census**, we measure the variable/s of interest from all the elements of the population.
- In **sampling**, we measure the variable/s of interest from the elements belonging in one subset of the population.

Advantages of Sampling over Complete Enumeration

- (1) more economical, (2) faster to accomplish, (3) wide scope of study is more viable, (4) measurement errors are more likely to be smaller, (5) sometimes complete enumeration is not feasible

Basic Concepts

Definition 3.1.

- The **target population** is the population we want to study.
- The **sampled population** is the population from where the sample is actually selected.

Remarks

- Inferences based on sample data will apply to the sampled population.
- Ideally, the target and the sampled populations should be the same.

Example

Target population:

Collection of all residents in Metro Manila

Suppose the sample of residents was selected from the list of names in the telephone directory [of PLDT].

Sampled population:

Basic Concepts

Definition 3.2.

- The **elementary unit** or **element** is a member of the population whose measurement on the variable of interest is what we wish to examine.
- The **sampling unit** is the unit of the population that we select in our sample.

Note: Before selecting the sample, *the population must first be divided into non-overlapping sampling units*. Sometimes the sampling unit is the element itself. Sometimes the sampling unit is a group of more than one element.

Example

A researcher wishes to estimate the total number of children below 12 years old in Barangay X. In order to do this, he selects a sample of households in Barangay X by selecting a sample of blocks in the barangay then collects data on the number of children below 12 years old from each household in the selected blocks.

1. What is the variable of interest?
2. What is the parameter of interest in this study?
3. What are the elementary units in this study?
4. What are the sampling units in this study?

Basic Concepts

Definition 3.3.

The **sampling frame** or **frame** is a list or map showing all the sampling units in the population.

Note: The researcher will select sample from the sampling frame. The sampling frame will define the sampled population.

Example

Suppose a researcher is interested in getting the opinion of eligible voters on the media campaign of candidates running for top positions in the government.

Target population: set of all eligible voters

Sampling frame: Commission on Elections (COMELEC) list of registered voters

Sampled population: set of registered voters in the list of COMELEC

Basic Concepts

Definition 3.4

Sampling error is the error attributed to the variation present among the computed values of the statistic from the different possible samples consisting of n elements.

Non-Sampling error is the error from other sources apart from sampling fluctuation. (example: measurement errors)

Example

Suppose we conducted a census on a small population consisting of N=15 students. From each student, we measured the variable of interest, **X=weekly allowance**.

400 400 450 475 500 500 500 525

550 575 600 700 750 750 800

Let the parameter of interest be the total weekly allowance of all students in the population.

Total = 400 + 400 + 450 + 475 + 500 + 500 + 500 + 525 + 550 + 575 + 600 + 700 + 750 + 750 + 800 = **8,475 pesos**

Example

Let us take a sample of $n = 5$ students using systematic sampling.

Sample	Sample Data
1	400 475 500 575 750
2	400 500 525 600 750
3	450 500 550 700 800

$$\boldsymbol{estimated\ total} = \left(\frac{N}{n}\right) x\ (sample\ total) = \left(\frac{15}{5}\right) x\ (sample\ total) = 3\ x\ (sample\ total)$$

Sample	Computation	Estimated Total
1	3 x (400 + 475 + 500 +575 + 750)	8,100
2	3 x (400 + 500 +525 + 600 + 750)	8,325
3	3 x (450 + 500 + 550 + 700 + 800)	9,000

Basic Concepts

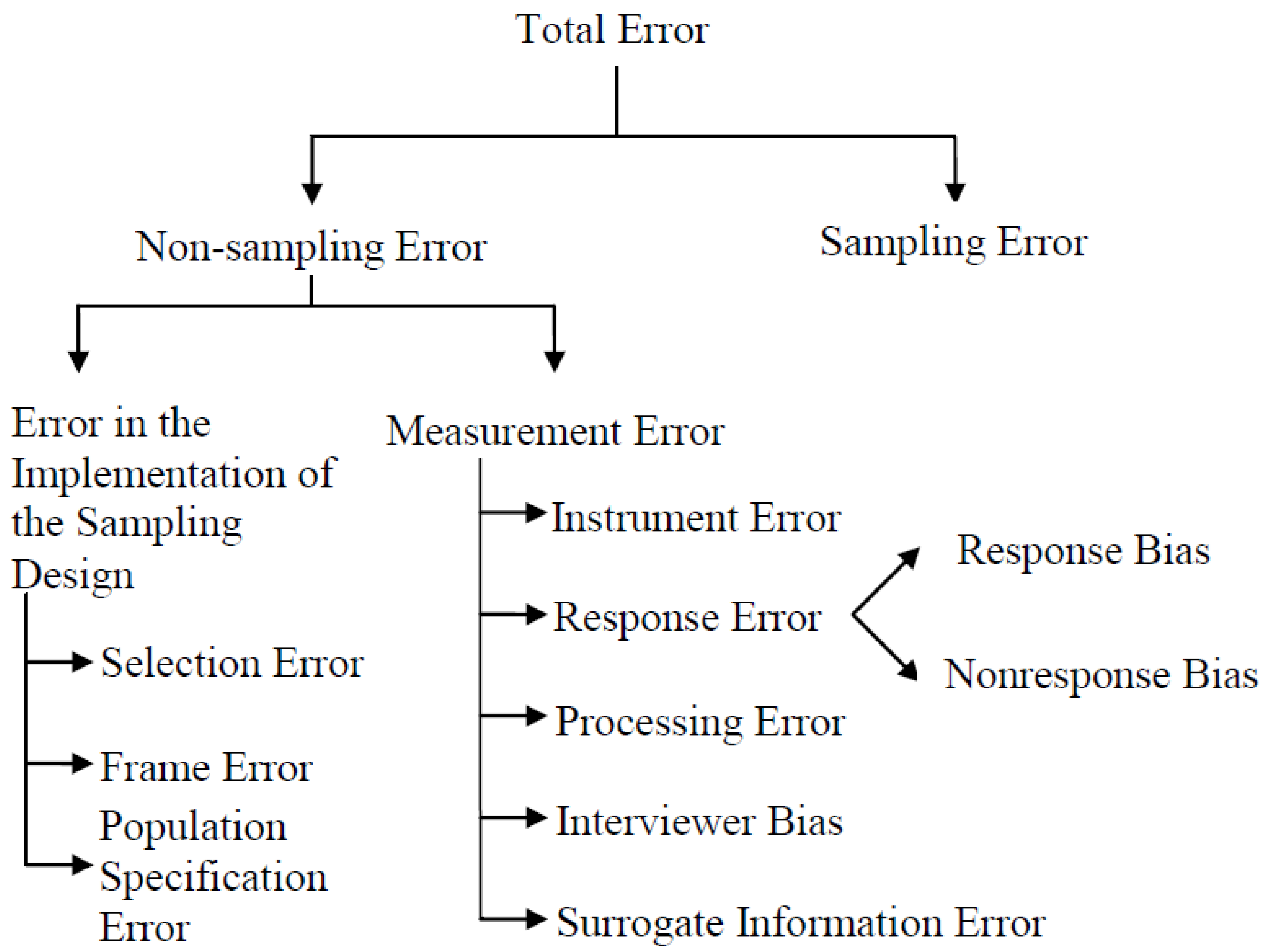
Notes:

To manage the possible sampling error, one must have a good sampling design

- Choose an appropriate sample selection procedure.
- Sample size must be large enough.
- Select the right statistic to estimate the parameter of interest.

Census results **do not** have sampling errors

Basic Concepts



Total Error

Non-sampling Error

Sampling Error

Error in the Implementation of the Sampling Design

Selection Error

Frame Error

Population Specification Error

Measurement Error

Instrument Error

Response Error

Response Bias

Nonresponse Bias

Processing Error

Interviewer Bias

Surrogate Information Error

Two Methods of Sampling

Definition 3.5.

Probability Sampling: method of selecting a sample wherein each element of the population has a known, nonzero chance of being included in the sample.

Non-probability sampling: method of selecting a sample wherein some elements of the population do not have a known chance of inclusion in the sample or the chance is zero.

Remarks

In probability sampling, the **chances of inclusion need not be the same for all elements**.

The use of a **randomization mechanism** in the selection of the sampling units included in the sample makes it possible for us to **determine the chances of inclusion**.

Knowing the chances of inclusion will allow us to **determine** if our estimates are **reliable** and **accurate**.

<u>Question</u>: Will the use of probability sampling assure us of selecting a “representative sample”?

NO!

Methods of Probability Sampling

We present some methods of probability sampling:

- Simple random sampling (SRS)
- Stratified sampling
- Systematic sampling
- Cluster sampling
- Multi-stage sampling

Simple Random Sampling

Definition 3.6.

Simple Random Sampling (SRS) is a probability sampling method wherein all possible subsets consisting of n elements selected from the N elements of the population have the same chances of selection.

Note: Consequently, all the elements of the population have the same chances of inclusion in the sample.

Simple Random Sampling

Two Types of SRS:

Simple random sampling without replacement (SRSWOR): all the n elements must be distinct

Simple random sampling with replacement (SRSWR): the n elements need not be distinct

Sample Selection Procedure: SRS

Step 1. Look for a sampling frame listing down all the elements in the population. Assign a unique serial number, from 1 to N, to each one of the elements in the list.

Step 2. Generate n numbers from 1 to N using any randomization mechanism.

- These n numbers must be distinct if we are using simple random sampling **without** replacement.
- In simple random sampling **with** replacement, these n numbers need not be distinct.

Step 3. The sample will consist of the elements with the same serial number as the numbers generated in Step 2.

Example: SRSWOR

Population = {Janine, Josiel, Jan, Eryll, and Eariel}. Select a sample of size 2 using SRSWOR.

There are 10 possible subsets of size 2 containing distinct elements selected from the 5 elements in the population as follows:

{Janine, Josiel}
{Janine, Jan}
{Janine, Eryll}
{Janine, Eariel}
{Josiel, Jan}
{Josiel, Eryll}
{Josiel, Eariel}
{Jan, Eryll}
{Jan, Eariel}
{Eryll, Eariel}

By definition, all 10 samples have the same chances of selection.

All 5 elements of the population have the same chances of inclusion. The common chance is 4/10=2/5 or 0.4. In general, it is $\frac{\boldsymbol{n}}{\boldsymbol{N}}$.

Example: SRSWR

Population = {Janine, Josiel, Jan, Eryll, and Eariel}. Select a sample of size 2 using SRSWR.

There are 25 possible ordered samples of size 2, where the coordinates need not be distinct.

(Janine, Janine)	(Janine, Josiel)	(Janine, Jan)	(Janine, Eryll)
(Janine, Eariel)	(Josiel, Janine)	(Josiel, Josiel)	(Josiel, Jan)
(Josiel, Eryll)	(Josiel, Eariel)	(Jan, Janine)	(Jan, Josiel)
(Jan, Jan)	(Jan, Eryll)	(Jan, Eariel)	(Eryll, Janine)
(Eryll, Josiel)	(Eryll, Jan)	(Eryll, Eryll)	(Eryll, Eariel)
(Eariel, Janine)	(Eariel, Josiel)	(Eariel, Jan)	(Eariel, Eryll)
(Eariel, Eariel)			

By definition, all 25 ordered samples have the same chances of selection.

All 5 elements have the same chances of inclusion This common probability is 9/25=1 – $4^2/5^2$. In general, it is $\mathbf{1-\frac{(N-1)^n}{N^n}}$.

Simple Random Sampling

Advantages

- Design is simple and easy to understand
- Estimation methods are simple and easy

Disadvantages

- It needs a list of all elements in the population
- Sample size must be very large for heterogeneous populations in order to get reliable results
- High transportation cost if elements are widely spread geographically

Simple Random Sampling

When to use:

- If the elements are **homogeneous** with respect to the characteristic under study
- If the elements are not so spread out geographically (why?)

Stratified Sampling

Definition 3.7.

Stratified sampling is a probability sampling method where we divide the population into nonoverlapping subpopulations or strata, and then select one sample from each stratum. The sample consists of all the samples in the different strata.

Note:

If the sample selection procedure is SRSWOR for all of the strata, then we specifically refer to this method as **stratified random sampling**.

Sample Selection Procedure

Step 1. Identify the variable whose categories will serve as the strata in the study. This will be the **stratification variable**.

Step 2. Look for a sampling frame that lists down all of the elements in the population and contains data on the value of the stratification variable for each element.

Step 3. Use the data on the stratification variable to place each element of the population into its appropriate stratum.

Step 4. Select a sample from each stratum using SRSWOR.

Step 5. The sample will consist of all the samples selected in the different strata.

Stratification Variable

- In stratified random sampling, we get estimates that are more reliable if the stratification variable partitions the population into strata wherein the **elements within the same stratum are homogeneous with respect to the characteristic of interest**, but the strata themselves vary considerably among each other.
- Choice of strata may facilitate the administration and supervision of data collection (ex: geographic subdivision)
- Strata may be the subpopulations of interest
- Information on stratification variable must be available for each element of the population

Stratified Random Sampling: Example

Variable of interest: farm production

Parameter of interest: total farm production

Stratification variable: size of farm with categories small, medium, and large

How does this compare with SRS?

Stratified Random Sampling: Example

Identify an appropriate stratification variable for the following variables under study:

1. Height of schoolchildren in an elementary school
2. Annual earnings of UP graduates since 1990

Allocation of Sample

- Involves determining the sample size for each stratum
- Affects the chance of including an element in the sample

Proportional allocation

- The only allocation method wherein each element is given the same chance of selection. However, the sampling error will not always be expected to be small when we use this method.
- Under proportional allocation, the size of the sample in each stratum is proportional to the population size of the stratum.

Proportional Allocation

Example 3.14.

Suppose we want to get the opinion of business administration college students regarding premarital sex. A good stratification variable is sex views because the of the males may be very different from the views of the females. The population consists of $N = 500$ business administration students and the sample size is $n = 50$. Out of the 500, there are 300 female and 200 male students. The list of business administration students, together with their respective sex, is available at the records section of the college, or at the Office of the Registrar.

Σ Proportional Allocation

Stratum Number	Sex	Population Size	Proportion of Students	Sample Size
1	Male	N_1=200	200/500=0.4	n_1=50x0.4=20
2	Female	N_2=300	300/500=0.6	n_2=50x0.6=30
		N=500		n=50

Systematic Sampling

Definition 3.8.

Systematic sampling is a probability sampling method wherein the selection of the first element is at random and the selection of the other elements in the sample is systematic by subsequently taking every kth element from the random start, where k is the **sampling interval**.

SSP : Systematic Sampling Method A

Use this method when n is a divisor of N

Step 1: Decide on a method of assigning a unique serial number, from 1 to N, to each one of the elements in the population.

Step 2: Compute the sampling interval, $k = N/n$.

Step 3: Select a number from 1 to k using a randomization mechanism. Denote the selected number by r. The element assigned to this number is the first element of the sample.

Step 4: The other elements of the sample are those assigned to the numbers, $r + k, r + 2k, r + 3k, \ldots$ and so on until you get a sample of size n.

SSP : Systematic Sampling Method B

Use this method when n is NOT a divisor of N

Step 1: Treat the list as a circular list.

Step 2: Compute for k as the greatest integer of N/n.

Step 3: Select a number from 1 to N using a randomization mechanism. Denote the selected number by r. The element assigned to this number is the first element in the sample.

Step 4: The other elements are those assigned to the numbers $r + k, r + 2k, r + 3k,$ ..., and so on until you get a sample of size n.

Systematic Sampling: Example

(3.15) Suppose we wish to conduct a survey on the opinions of senior Statistics majors on the computerized registration system. We can get a list of seniors from the Office of the College Secretary. This will serve as the sampling frame. If the list contains the names of N=50 seniors arranged alphabetically then we can select a sample of n=10 students using systematic sampling method A.

(3.16) Let us consider the same problem described in example 3.15. However, this time let us select a sample of n = 8 seniors from our population of N = 50.

Cluster Sampling

Definition 3.9.

Cluster sampling is a probability sampling method wherein we divide the population into nonoverlapping groups or clusters consisting of one or more elements, and then select a sample of clusters. The sample will consist of all the elements in the selected clusters.

<u>Notes</u>:

The **sampling units** in cluster sampling are the **clusters** and not the elements.

If the clusters are selected using SRSWOR, then this is called **simple one-stage cluster sampling**.

SSP : Simple One-Stage Cluster Sampling

Step 1: Divide the population into nonoverlapping clusters.

Step 2: Number the clusters in the population from 1 to N.

Step 3: Select n distinct numbers from 1 to N using a randomization mechanism. The selected clusters are the clusters associated with the selected numbers.

Step 4: The sample will consist of all the elements in the selected clusters.

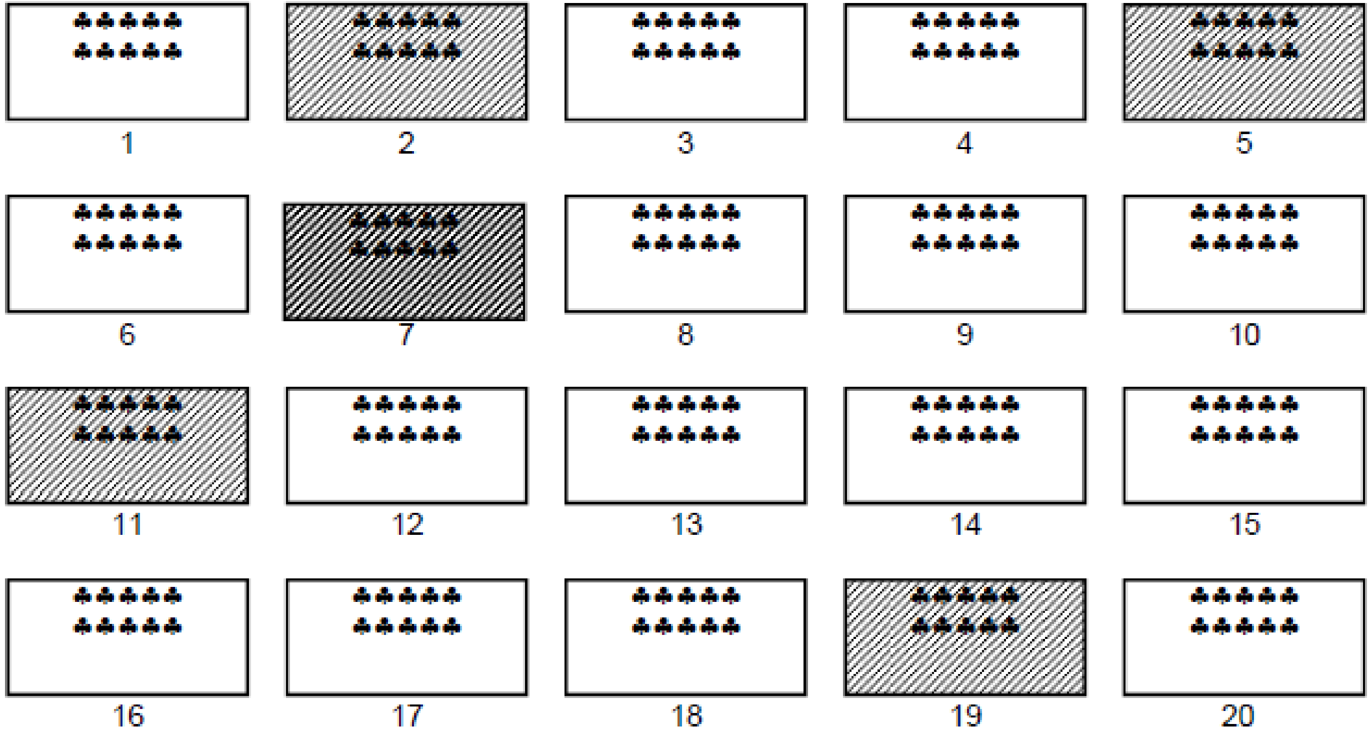
Example

Clusters: sections consisting of 10 students

(How many sections are there?)

Elements: students

(How many students per section? In total?)



Examples of Clusters

- city blocks serve as clusters of residents
- hospitals serve as clusters of patients
- schools serve as clusters of teachers
- class sections serve as clusters of students
- elements other than people are often sampled in clusters.
    - A car forms a cluster of four tires for studies of tire wear and safety.
    - A circuit board manufactured for a computer forms a cluster of semiconductors for testing.
    - A mango tree forms a cluster of mangoes for investigating insect infestation.
    - A plot in a forest contains a cluster of trees for estimating timber volume or proportions of diseased trees.

Stratified Sampling vs Cluster Sampling

Stratified Sampling	Cluster Sampling
• A sample is selected from each stratum. Consequently, all strata will be represented in the sample. • Estimates will be more reliable if the elements in a stratum are homogeneous with respect to the characteristic under study	• Only a sample of clusters will be selected in the sample then all elements in the selected clusters will be included in the sample • Estimates will be more reliable if the elements in a cluster are heterogeneous with respect to the characteristic under study

Multi-Stage Sampling

Definition 3.10

Multistage sampling is a probability sampling method where there is a hierarchical configuration of sampling units and we select a sample of these units in stages.

Multi-Stage Sampling

In the first stage of sampling, the sampling units are called **primary stage units** (PSUs). In the second stage of sampling, the sampling units are called **second stage units** (SSUs). In the third stage of sampling, the sampling units are called **third stage units** (TSUs). And so on.

The **sample selection procedure may vary in each stage**. If the elements in the sample can already be identified on the second stage and the sample selection procedure is SRSWOR in each stage then this is referred to as simple 2-stage sampling. If the elements in the sample can already be identified on the third stage and SRSWOR is used in each stage then this is referred to as simple 3-stage sampling

SSP : Simple 3-Stage Sampling

Step 1. List the PSUs in the population and number them from 1 to N.

Step 2. Select a sample of PSUs using SRS.

Step 3. For each of the selected PSUs, list down the SSUs and number them from 1 to the total number of SSUs.

Step 4. Select a sample of SSUs from each one of the selected PSUs using SRS.

Step 5. For each of the selected SSUs, list down the TSUs and number them from 1 to the total number of TSUs. The TSUs are already the elementary units in the study.

Step 6. Select a sample of TSUs from each one of the selected SSUs using SRS.

Step 7. The sample consists of all the TSUs selected in step 6.

Examples

Suppose we wish to study the expenditure patterns of households in the National Capital Region (NCR). We can select a sample of households for this study using simple three-stage sampling where the PSUs are the cities/municipalities, the SSUs are the barangays, and the third-stage units TSUs are the households.

Σ Exercise

Identify the probability sample selection procedure used in each of the following cases:

a) A survey obtained a sample of laborers by first classifying the different areas as either rural or urban. After which, a sample of laborers was taken from each area.

b) To select a sample of households in the country, a sample of provinces were selected, then a sample of municipalities were chosen from each of the selected provinces, then a sample of barangays were chosen from each of the selected municipality, and all households in the selected barangays were included.

c) A study selected a sample of municipalities from every province in the country and included all child laborers in the selected municipalities.

d) In the game of lotto, they select six balls from a container with 42 balls.

e) A car manufacturer conducts quality checking on every 20th car in the production line from the random start.

Non-probability Sampling

Accidental/Convenience sampling

Items or units that are most accessible are included in the sample.

(Examples include a sample of volunteers, street corner interviews, and pull-out questionnaires in a magazine.)

Non-probability Sampling

Purposive sampling

Sample is selected in accordance with an expert's subjective judgment on what a "representative" sample should contain. Usually sampling units are clusters whose profile is similar to the profile of the population.

Example: identify barangays where the income distribution of households is the same as the population

Non-probability Sampling

Quota sampling

It is the nonprobability sampling version of stratified sampling where the researcher just sets a quota or number of sampling units in each grouping but uses convenience sampling in selecting the units in each group.

Example: sample includes any 50 Globe subscribers and any 50 Smart subscribers