# Correcting Common Misconceptions in Statistics: For K-12

**Josefina V. Almeda, PhD**
Executive Director III
Philippine Statistical Research and Training Institute
University of the Philippines Diliman (on secondment)

“The trouble with statistics is that they are often counter-intuitive. What seems like a common sense answer to a question is often wrong.”

**Daniel Ben-Ami**

**Research has shown that adults have intuitions about probability and statistics that, in many cases, are at odds with accepted theory**

Clifford, K.S

# BACKGROUND

- **Teacher:**
  - Share the statistical concepts in a symmetrical way
  - Follow a logical sequence to introduce statistical concepts
- **Student:**
  - May not pay attention to all the pieces
  - Learn fragmented parts
  - Partial understanding of the subject matter

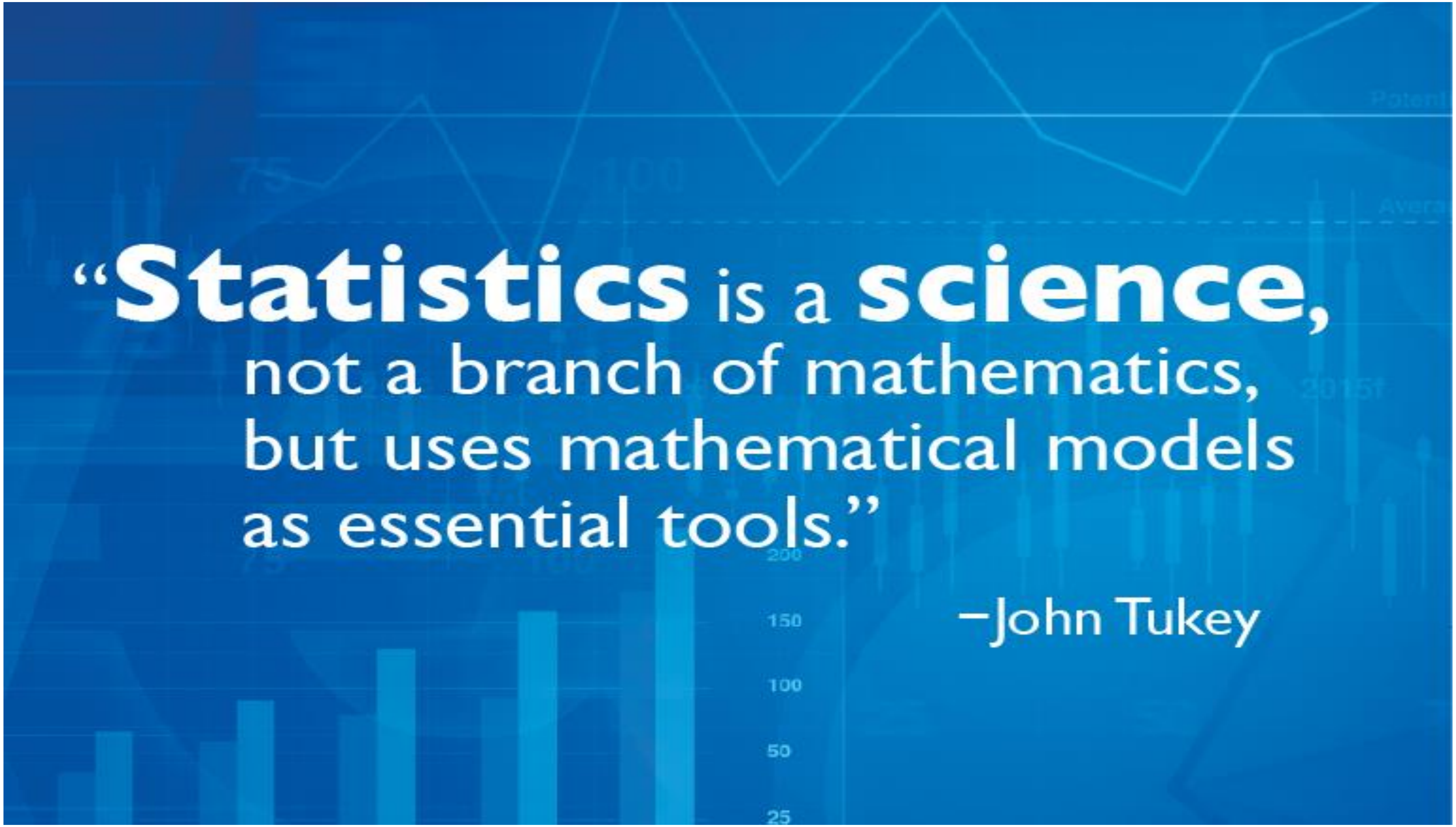
# MISCONCEPTION

- **Definition**: Misconception is a wrong belief or opinion as a result of not fully understanding something
- **Cause:** Some statistical concepts are interrelated – known portions to fill in the missing/unknown part – similar terminologies
- **Remedy:** Spot out the related components for reinforcement

# Misconceptions about Statistics

# Misconception about Statistics

- **Misconception:** Statistics is Math.
- Statistics uses numbers but numbers are not the primary focus of statistics, at least to most practitioners.



"**Statistics** is a **science**, not a branch of mathematics, but uses mathematical models as essential tools."

–John Tukey

You can love Statistics and be good at statistical thinking even if you think you hate Math.

# Misconception about Major Fields of Statistics

- **Misconception:** The two major fields of Statistics are Descriptive Statistics and Inferential Statistics

- Descriptive Statistics and Inferential Statistics are two major areas under Applied Statistics

# Two Major Fields of Statistics

## 1. Applied Statistics

- Concerned with the procedures and techniques used in the collection, presentation, organization, analysis, and interpretation of data.

## 2. Theoretical or Mathematical Statistics

- Concerned with the development of the mathematical foundations of the methods used in applied statistics.

# Two Major Areas in Applied Statistics

## Descriptive statistics

- Includes all the techniques used in organizing, summarizing, and presenting the data on hand <u>without drawing conclusions</u> or inferences about a large group

## Inferential statistics

- Includes all the techniques used in analyzing the sample data that will lead to <u>generalizations</u> about a population from which the sample came from

# Misconception about Major Fields of Statistics

- **Misconception:** The value of the variable can be collected by asking the question WHAT.

- The value of the variable can be collected by asking the questions What, How, When, Whom, Where

# Misconception about Major Fields of Statistics

Recall:

- **Variable** is a characteristic or attribute of persons or objects which can assume different values for different persons or objects

# Misconceptions about Sources of Errors in Data Collection

# Misconception about Sampling Error

- **Misconception:** Sampling error is error in selecting a sample.

- Sampling error occurs when we collect data from a sample and not from all the elements in the population.

# Sources of Errors in Data Collection

- ***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 fluctuations.

# Example of Sampling Error

Suppose we conducted a census on a small population consisting of $N=15$ students.
Let $X$=weekly allowance.  The collected data set:

400 400 450 475 500 500 500 525
550 575 600 700 750 750 800

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

**Can we compute for the value of this parameter?**

# Example of Sampling Error

The computed value is:

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

Let us take a sample of *n*=5 students from the *N*=15 students in the population. Suppose we select our sample using *systematic sampling*.

# Example of Sampling Error

- For our population with N=15 students, there will only be 3 possible systematic samples, with 5 students each. The respective data sets for these 3 samples are as follows:

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

# Example of Sampling Error

Suppose we use the following formula to estimate the population total using the sample data:

*estimated 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 |

# Notes on Sampling Error

The variation in the estimated totals of the 3 possible systematic samples in the example demonstrates the sampling variation mentioned in the definition of ***sampling error***.

When we use a particular sample selection procedure to collect data, we already expect that there will be variation in the computed estimates using the different possible samples containing *n* elements.

# Notes on Sampling Error

The size of the sampling error will depend on how varied these estimates are from each other and how close the estimated values are to the parameter of interest.

The sampling error will be small if the disparity of the estimate from the parameter for each one of the possible samples is generally small; otherwise, it will be large.

# Notes on Sampling Error

When the sampling error is small then we can be confident that the results that we get from our sample will not be too different from the results from the other samples of the same size and even the results computed from the population itself.

# Notes on Sampling Error

We can reduce the_size of sampling error in our study if we choose the

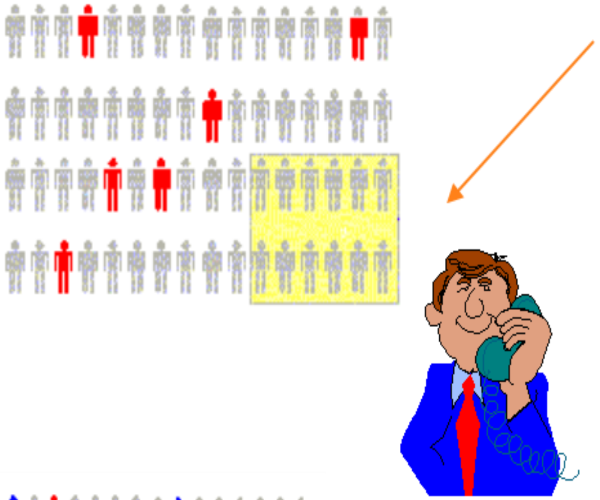
- appropriate sample selection procedure
- sample size and
- the formula to estimate the parameter of interest.

In short, we can reduce the sampling errors that we introduce in our results with a **good sampling design.**

# NonSampling Errors

## NonSampling Errors

- Coverage error
  - Appropriate Frame?
- Non response error
  - Follow ups?
- Measurement Error

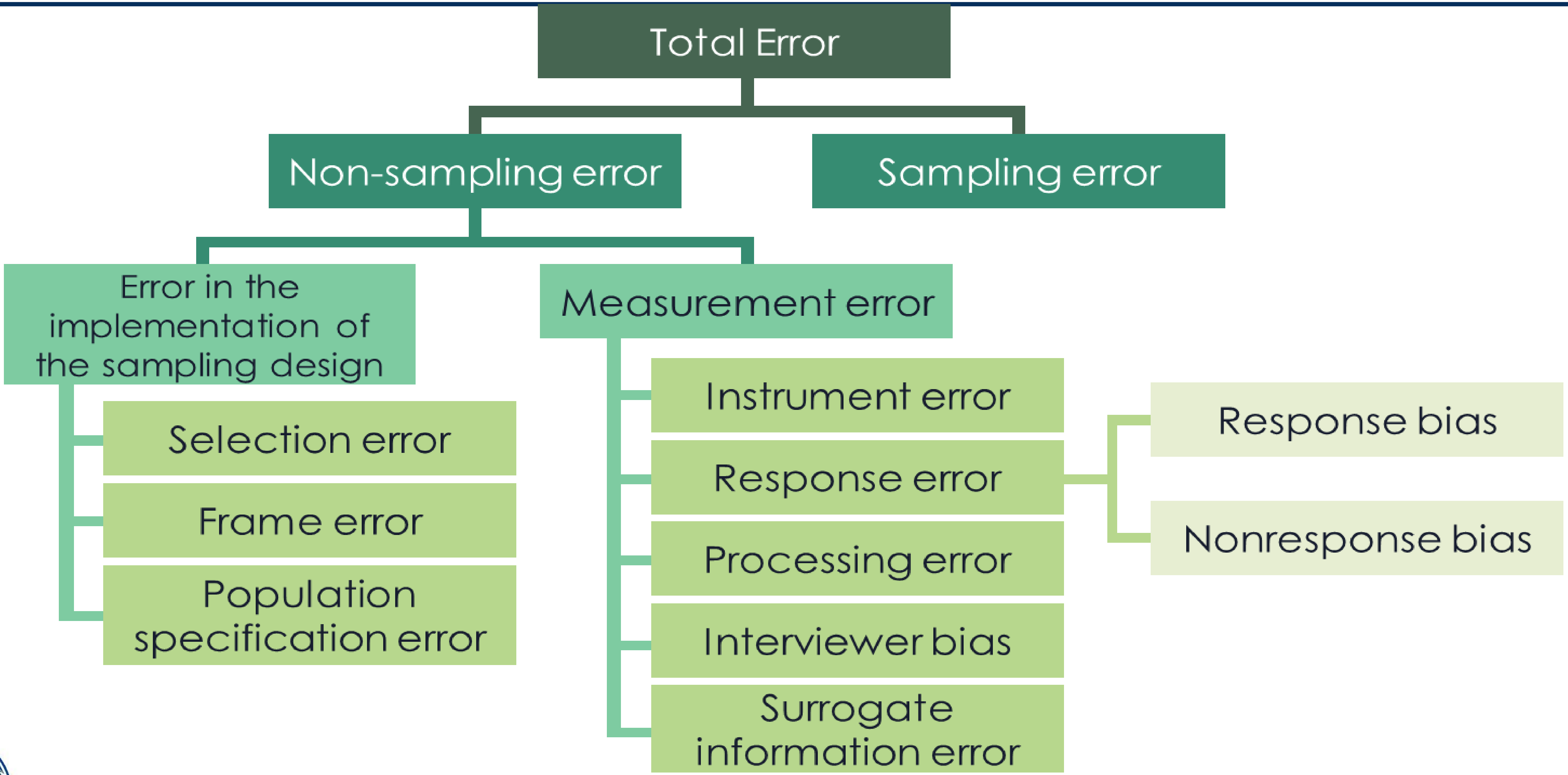


Excluded from frame.

Follow up on nonresponses.

Chance differences from sample to sample.

# Two Types of Errors



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

Processing error

Interviewer bias

Surrogate information error

Response bias

Nonresponse bias

# Questions in Sampling/NonSampling Errors

**Question No. 1:**

- Can we have sampling errors in a census?

**Question No. 2:**

- Can we have nonsampling errors in a census? How about sample survey?

**Question No. 3:**

- Which method of data collection can have more non sampling errors, sample survey or census?

# Questions in Sampling/NonSampling Errors

## Question No. 4:

- How can we minimize sampling errors?

## Question No. 5:

- How can we minimize nonsampling errors?

# Misconceptions about Sampling

# Misconception about Sampling

- **Misconception:** Sampling Unit and Element
- **Element** is an object on which a measurement is taken
- **Sampling Units** are nonoverlapping collections of elements from the population that cover the entire population

# Misconception about Sampling

Recall:

- **Population** is the a collection of all the elements under consideration in any statistical study

- **Sampling Frame** is a list of sampling units

# Misconception about Sampling

## Example of Element and Sampling Unit

Suppose we select our sample from a list of registered voters.

**Sampling unit** is a registered voter
**Element** is a registered voter

However, if we select our sample from a list of families in the barangay:
**Sampling unit** this time is a family, which may consist of one or more **elements** (registered voter) in the study.

# Misconception about Sampling

- **Misconception:** Sampled population is the sample of interest.
- Sampled population is the population from where we actually select the sample.

# Concepts in Sampling

The ***target population*** is the population we want to study.

The ***sampled population*** is the population from where we actually select the sample.

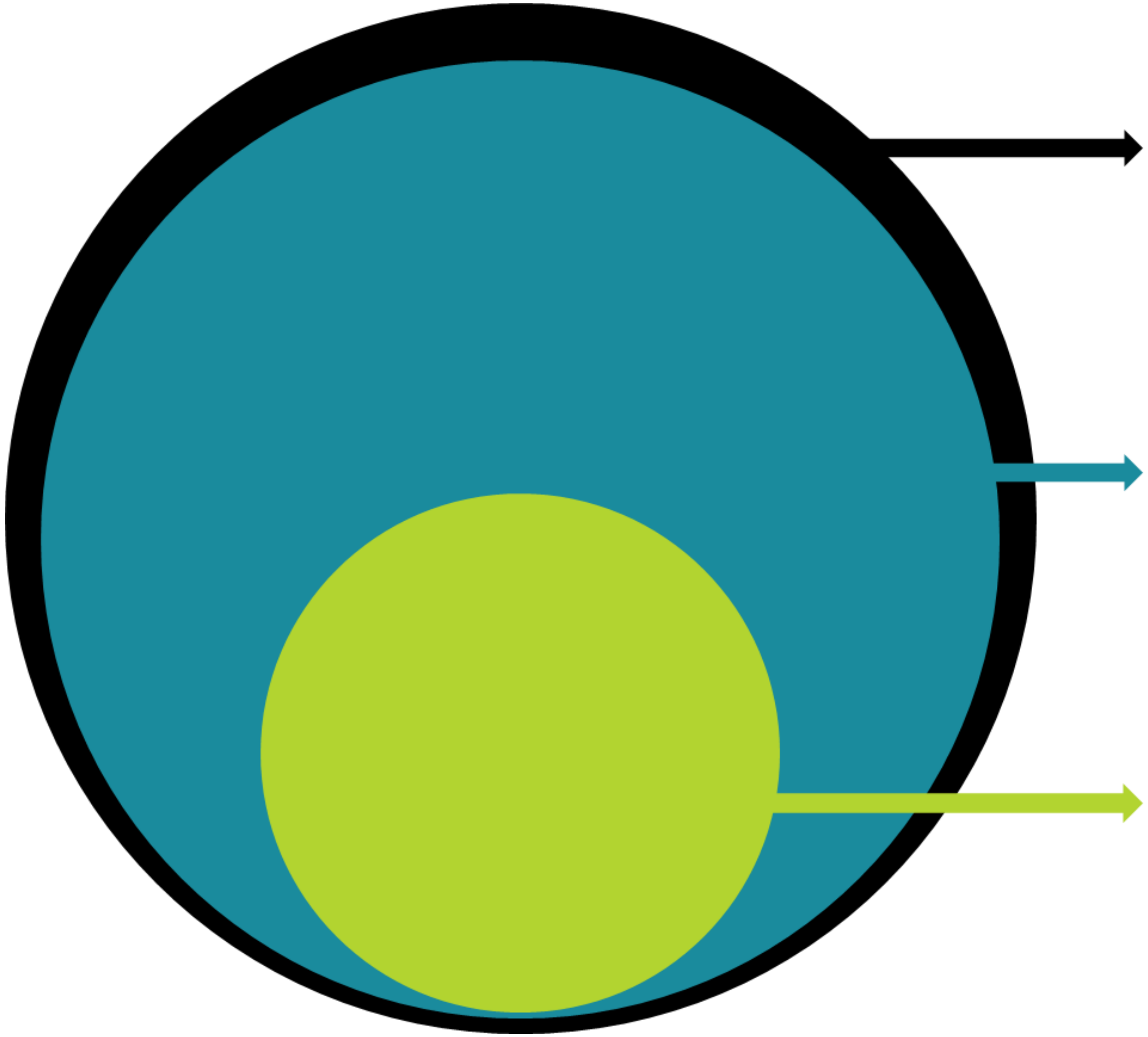
The ***sampling frame*** or ***frame*** is a list or map showing all the sampling units in the population.

# Concepts in Sampling

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

# Sampling and Sampling Techniques



Target population

Sampled population

Sample

# Concepts in Sampling

This sampled population excludes the eligible voters who did not register or did not revalidate their registration with the COMELEC during the registration period.

If COMELEC was not able to update this list completely then this sampled population will also include people who do not belong in the target population such as those who have passed away or who have changed their citizenship.

# Misconceptions about Sampling

- **Misconception:** It is not possible to come up with conclusions about a population consisting of millions of people simply by using sample data taken from only a few thousands of people.
- When the elements in the population are not too different from each other with respect to the variable under study, then there is no need to get data from each one of the elements in the population in order to describe its general characteristic.

# Misconception about Sampling

- **Misconception:** Probability sampling is widely believed as ‘a method of sampling so that each and every element of the population HAS THE SAME chance of being selected’.

- The correct definition is ‘each and every element of the population has a NON ZERO chance (not necessarily equal) of being selected’.
- The first definition only pertains to SIMPLE RANDOM SAMPLING, one of many possible methods of selecting a probability sample.

# 2 Methods of Sampling

**Probability Sampling** - procedure wherein every element of the population is given a (known) nonzero chance of being selected in the sample

**Nonprobability Sampling** - procedure wherein not all the elements in the population are given a chance of being included in the sample

# Methods of Probability Sampling

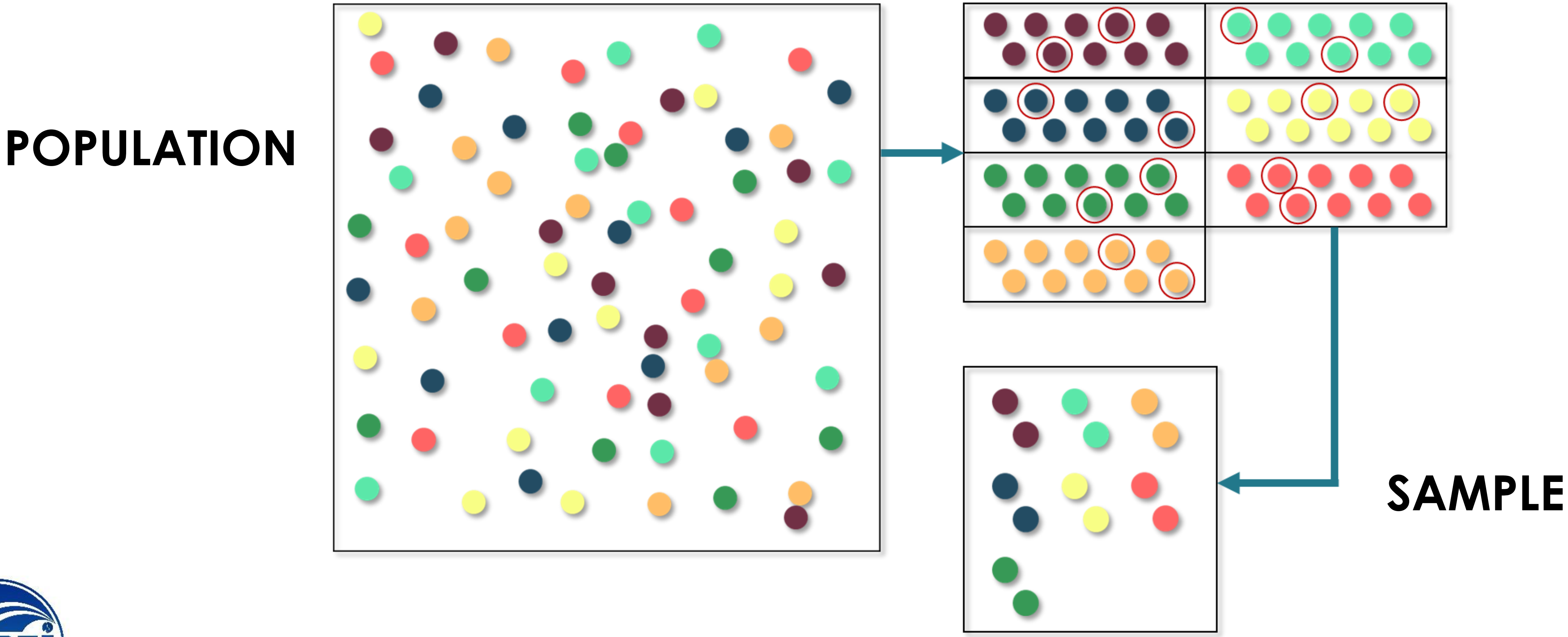
- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling
- Cluster Sampling
- Multistage Sampling

# Stratified and Cluster Sampling

- **Misconception:** Stratified and Cluster Sampling

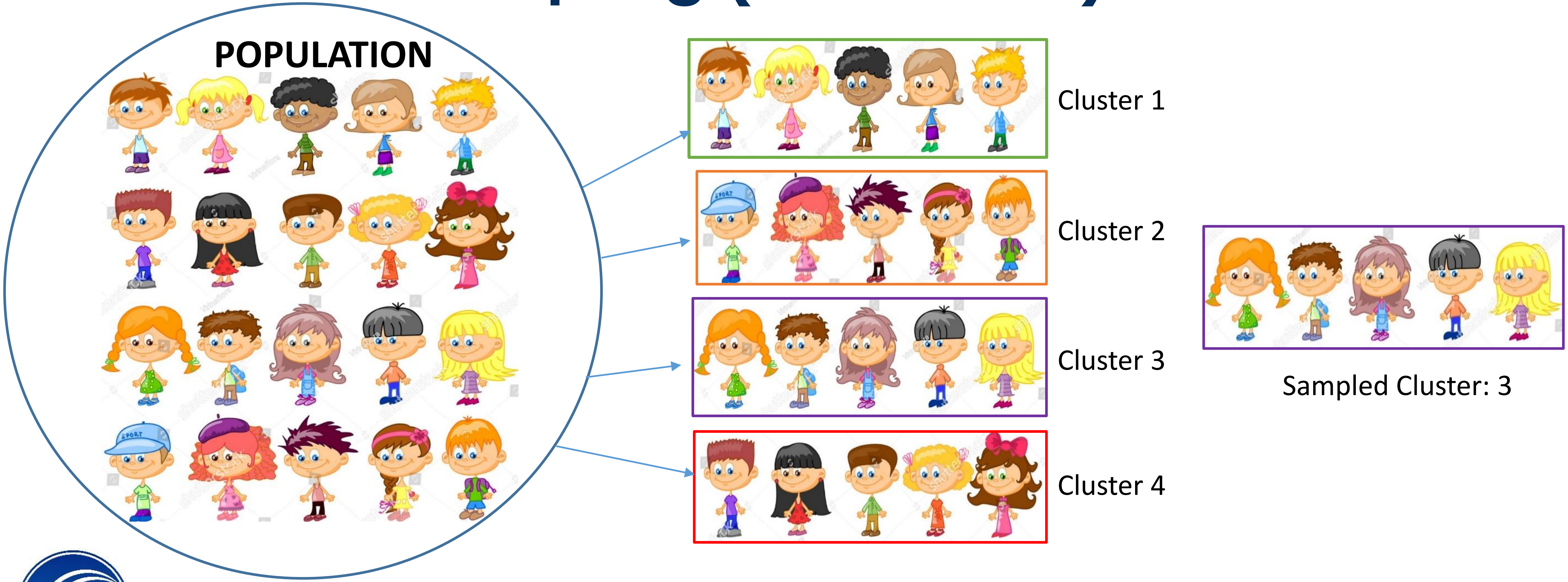
# Sampling and Sampling Techniques

## Stratified Sampling (Illustration)



# Sampling and Sampling Techniques

## Cluster Sampling (Illustration)



POPULATION

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Sampled Cluster: 3

# Methods of NonProbability Sampling

- Convenience Sampling
- Purposive Sampling
- Accidental Sampling
- Quota Sampling
- Volunteer Sampling

# Misconceptions in Sampling

- **Misconception:** The sample size should be determined as a fixed percentage of the size of the population (e.g., 10%, 20%).

- No theoretical basis. Sampling (proportion-wise) becomes more beneficial as the size of the population increases.
- Population size has very minimal effect on the determination of the sample size.

# Misconceptions in Sampling

**N = 100**

- 10% of N:  n =
- 20% of N:  n =
- 30% of N:  n =

**N = 100,000**

- 10% of N:  n =
- 20% of N:  n =
- 30% of N:  n =

# Misconceptions on Sampling

- **Misconception:** The 30 rule.
- This rule is primarily due to lessons learned in Central Limit Theorem.

Clarification: this rule only tells us that a sample size of at least 30 is considered “large” enough so that one can safely assume the normality of certain statistics – normality assumption simplifies the way we analyze data in particular when one test statistical hypotheses.

# Misconceptions on Sampling

- **Misconception:** Some people think that there is only one formula to compute for the sample size.

- The formula can be expressed in terms of cost, or of the desired level of reliability.

- There will be different formulas for the sample size for the different sample selection procedures, for the different parameters of interest and the different estimators used to infer on the parameter.

# Misconception on Sampling

- **Misconception:** Improper use of formulas, e.g., Slovin’s Formula

- Not proper to call it Slovin’s Formula since he did NOT derive it!
- Based on the formula, P=0.5 to get the largest possible sample.
- It is valid only under simple random sampling and any other design that is theoretically more efficient than simple random sampling (e.g., one-stage stratified sampling).

# Misconception on Sampling

- **Misconception:** Proportionate sampling is ALWAYS the best way to carry out stratified sampling.

- While proportionate sampling has desirable traits, it is considered one of the best if the interest is to estimate parameters pertaining to the entire population (descriptive purpose).

# Misconception on Sampling

- **Misconception:** Proportionate sampling is ALWAYS the best way to carry out stratified sampling.

- However, another usual purpose of stratification is to compare parameters between strata (analytic purpose). In such situations, ensure that the sample size in each stratum is large enough, which is not always the case with proportionate sampling.
- Proposed scheme: equal allocation. In the case of equal allocation, each element of the population has a different chance of being selected.

# Notes on Sample Size

- The number of elements that you include in the sample must not be too small because this will not allow you to come up with reliable estimates.
- Likewise, the number of elements in the sample must not be too large because this will only be a waste of money.
- Every additional observation in the study will add to the cost of the study.
- If the observation does not provide any additional information then it will be a waste of money to get that observation.

# Misconceptions in Survey Operations

# Misconception on Survey Operations

- **Misconception:** If a respondent cannot understand the question, the data collector may explain the question.

- The data collector may only repeat the question.
- If the question is not understood, this means the questionnaire did not undergo enough pretesting.

# Misconception on Survey Operations

- **Misconception:** The data collector may translate the English questionnaire to the dialect for the respondent to easily understand the question.

- Questionnaire should be translated to the dialect by a professional.
- After translating it to the dialect, another professional should translate it back to English to make sure that the two instruments are identical.

# Misconception on Survey Operations

- **Misconception:** Pretesting of the questionnaire is conducted only once.

- Several pretests need to be conducted if there are still items in the questionnaire that are not easily understood, too wordy, etc.

# Misconceptions on Graphical Presentation

# Misconception on Graphical Presentation

- **Misconception:** Scale figures on the vertical axis may be cut and may start with any number for both line chart and vertical bar chart.

- Scale figure of the vertical axis should always start with zero in presenting data in line and vertical bar charts.

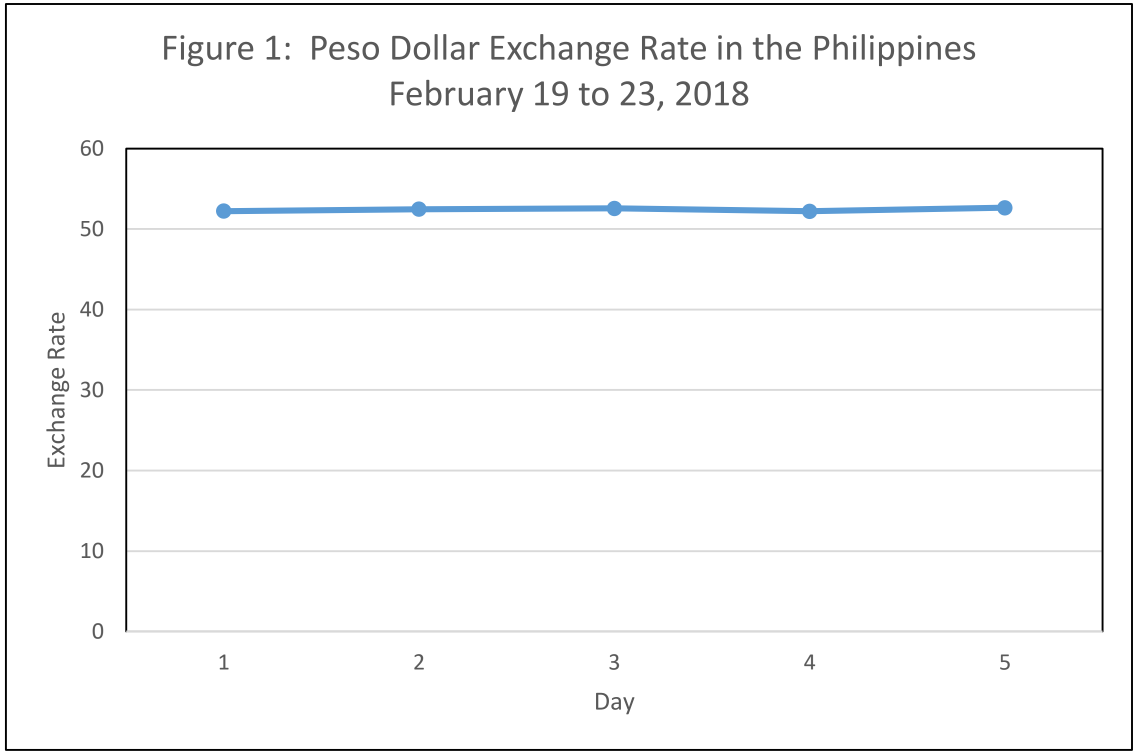


Figure 1: Peso Dollar Exchange Rate in the Philippines
February 19 to 23, 2018

With Zero on the Vertical Axis

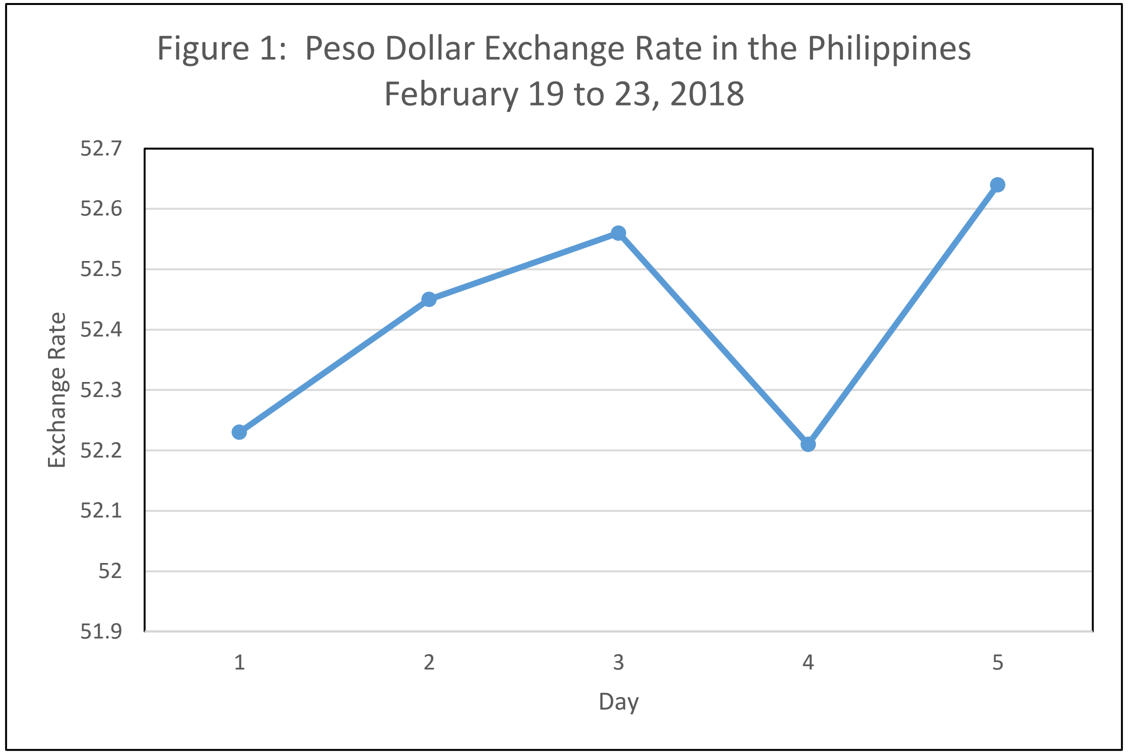


Figure 1: Peso Dollar Exchange Rate in the Philippines
February 19 to 23, 2018

Without Zero on the Vertical Axis

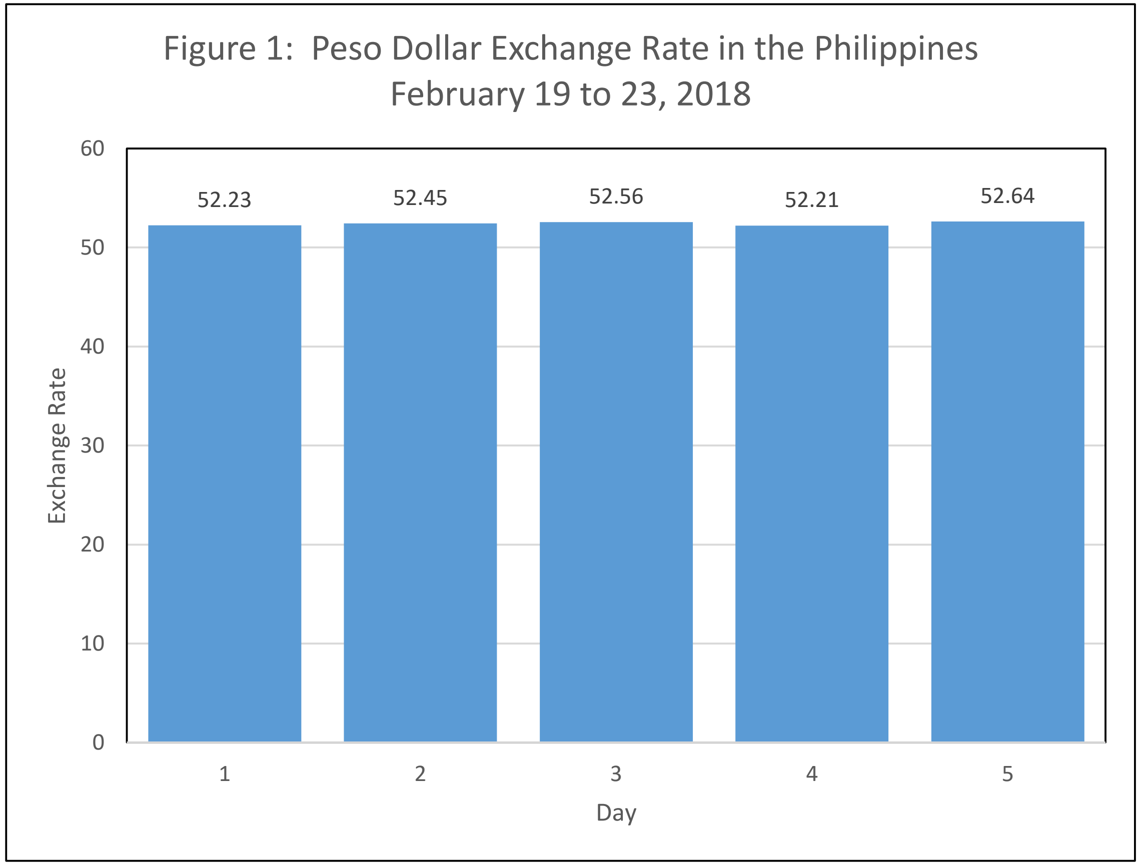


Figure 1: Peso Dollar Exchange Rate in the Philippines
February 19 to 23, 2018

With Zero on the Vertical Axis

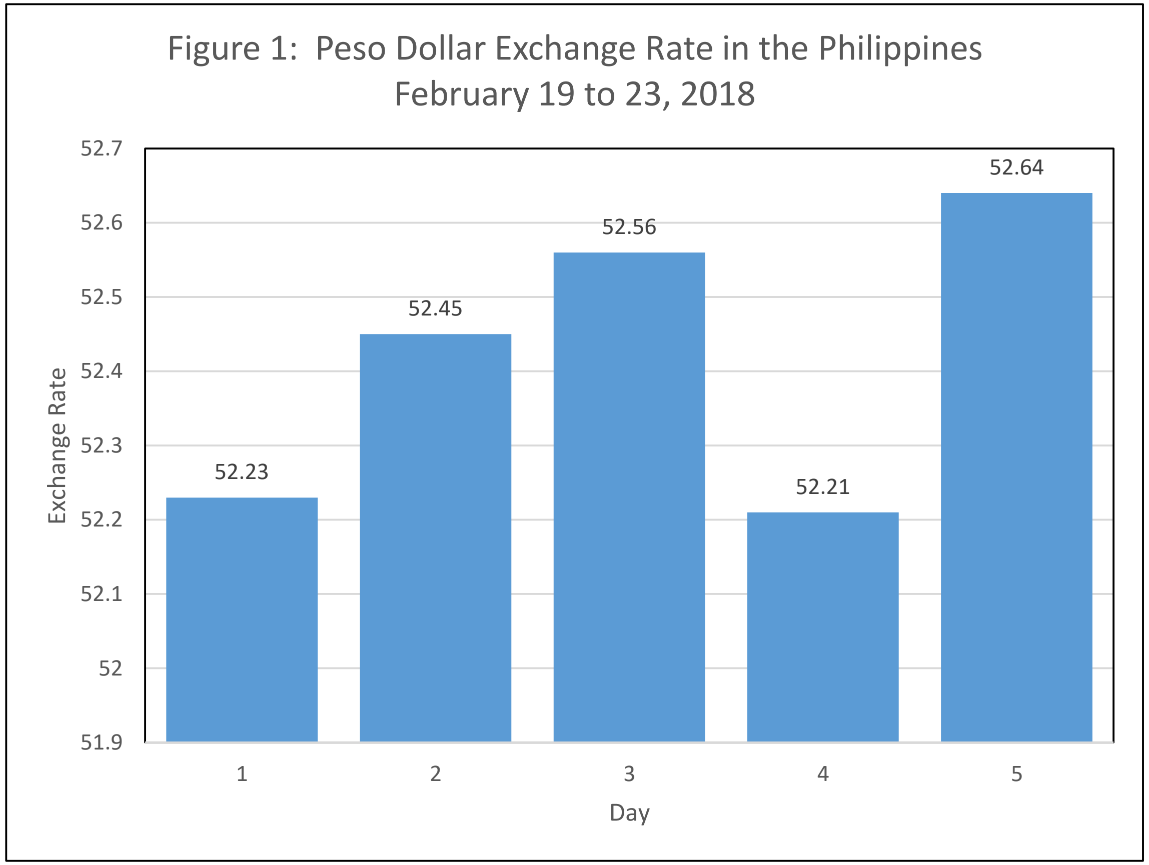


Figure 1: Peso Dollar Exchange Rate in the Philippines
February 19 to 23, 2018

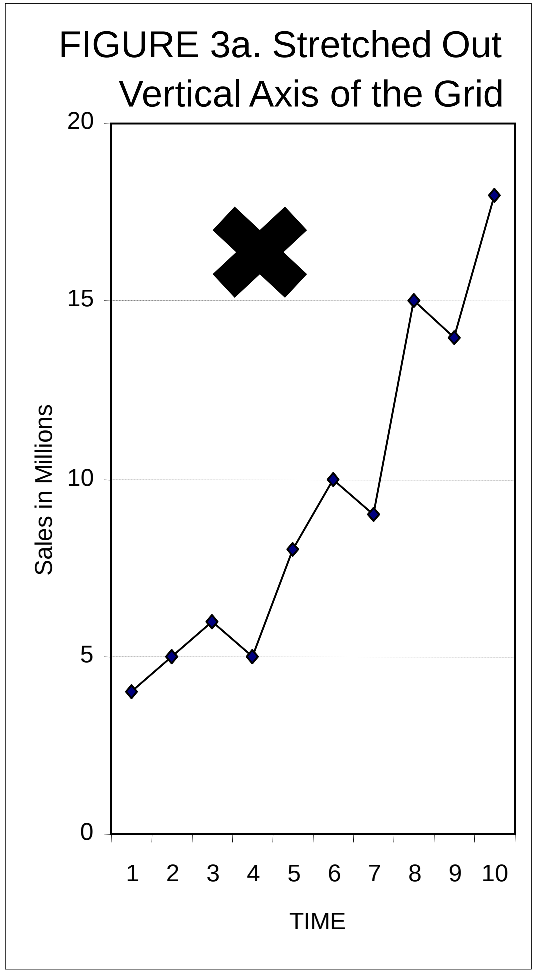
Without Zero on the Vertical Axis

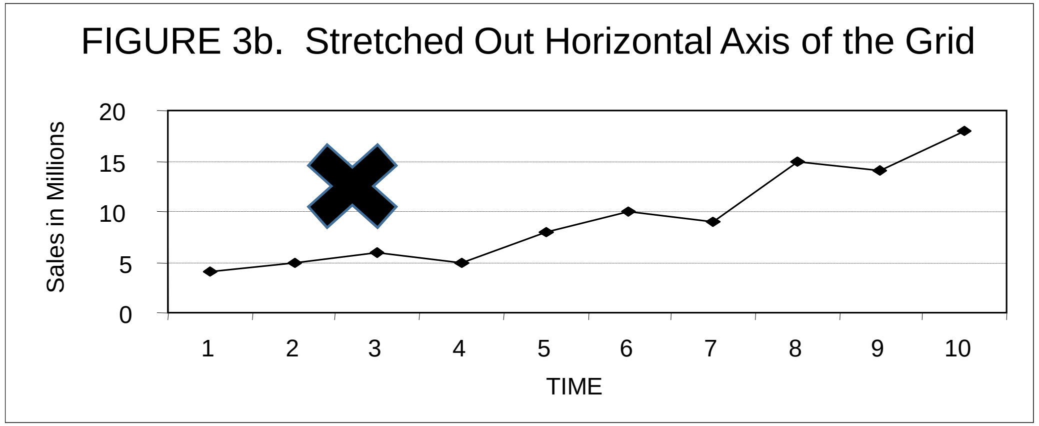
# Misconception on Graphical Presentation

- **Misconception:** There are no rules with regards to the size of the graph for the line chart.

- The ratio of height to width for the line chart should be 2:3 or 3:4.

# Methods of Data Presentation:  Line Chart





**FIGUREs 3a and 3b.  Stretched Out Vertical and Horizontal Axes and Its Consequences**

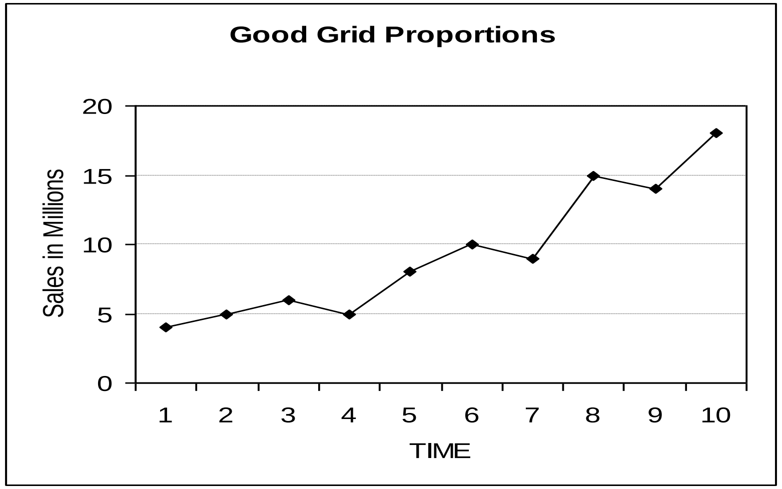


**FIGURE 3c.  Correct ratio of height to width (2:3)**

# Misconception on Graphical Presentation

- **Misconception:** We may use different colors for the vertical bar chart (column chart).

- Use only one color for the different bars of the column chart since the data is time series and there is only one variable of interest.

# Methods of Data Presentation: Column Chart

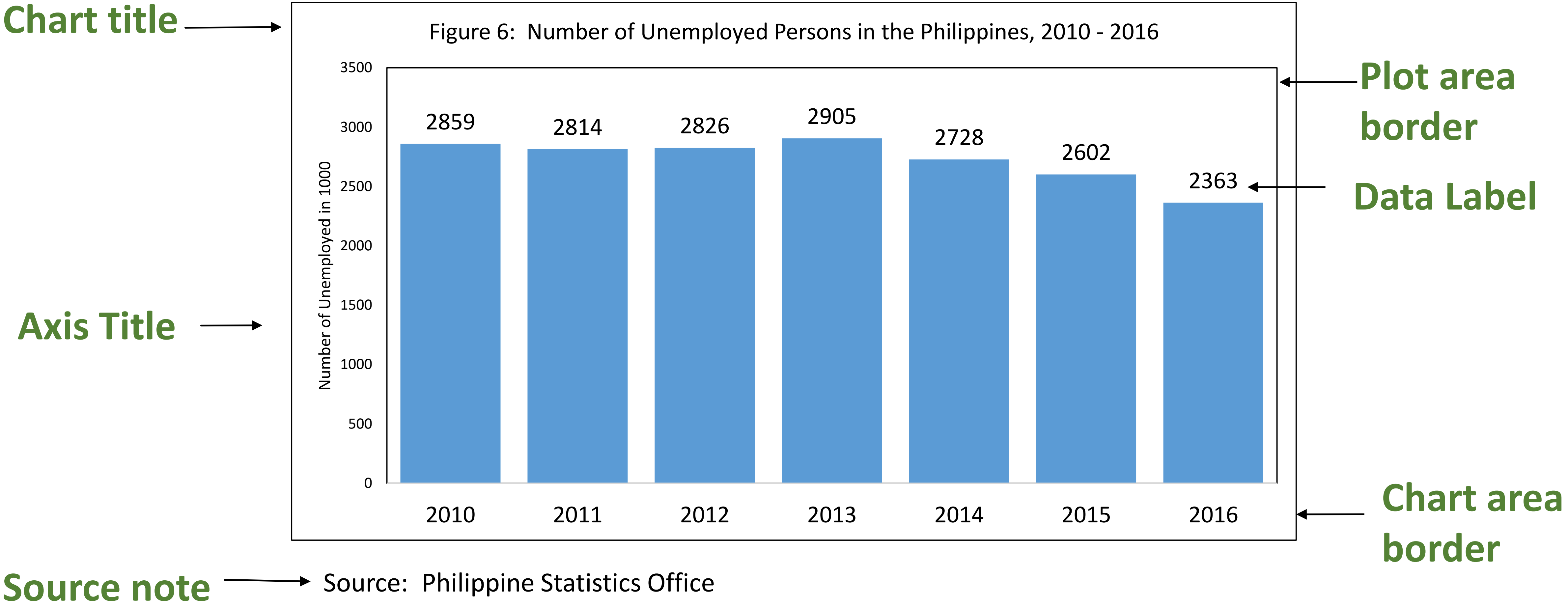


**Figure 6. Different Parts of a Column Chart**

# Misconception on Graphical Presentation

- **Misconception:** We use only one color for the different bars of the horizontal bar chart.
- We may use different colors for the different bars since the bars are independent of each other. But, do not overdo. One color is acceptable.

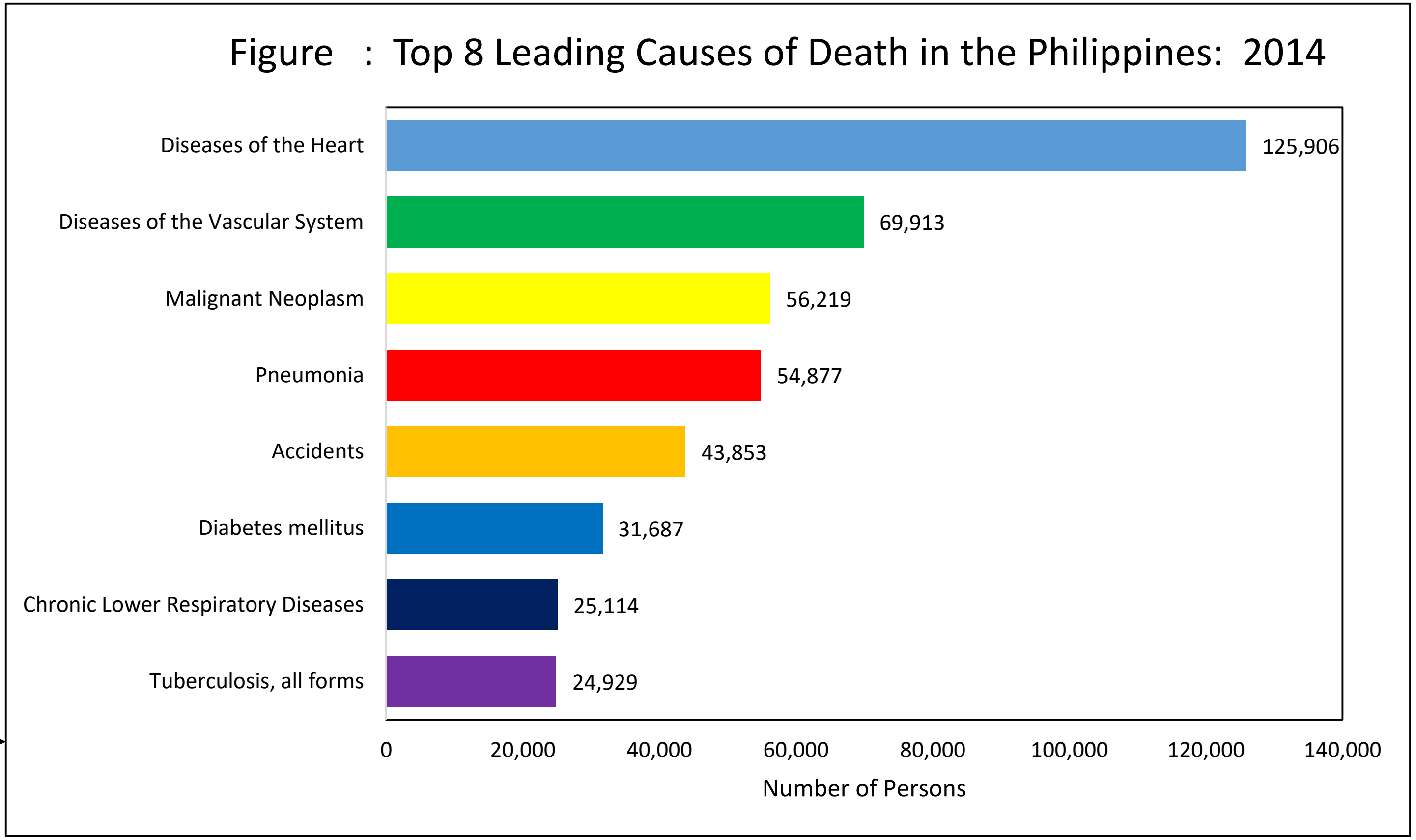
# Methods of Data Presentation

**Chart title**

**Axis Title**

**Scale figure**

**Source note**



Source: Department of Health

Figure 10. Different Parts of a Simple Horizontal Bar Chart

# Methods of Data Presentation

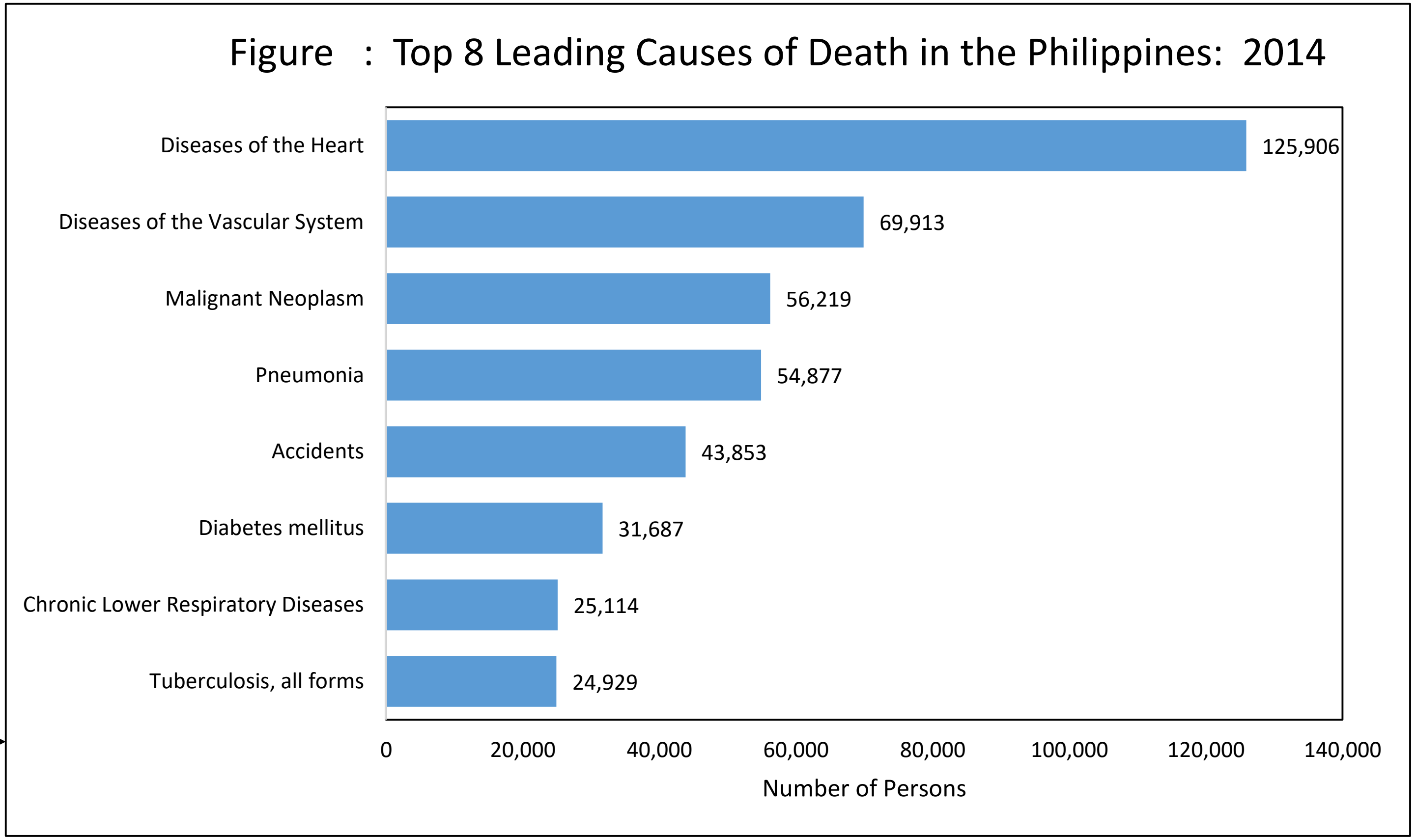


**Chart title**

**Axis Title**

**Scale figure**

**Source note**

Source: Department of Health

Figure 10. Different Parts of a Simple Horizontal Bar Chart

# Misconception on Graphical Presentation

- **Misconception:** The bars may be arranged according to magnitude including the Others category.

- The Others category even if it has a bigger percentage than the other categories should be placed last. The Others category is not included in the arrangement according to magnitude.

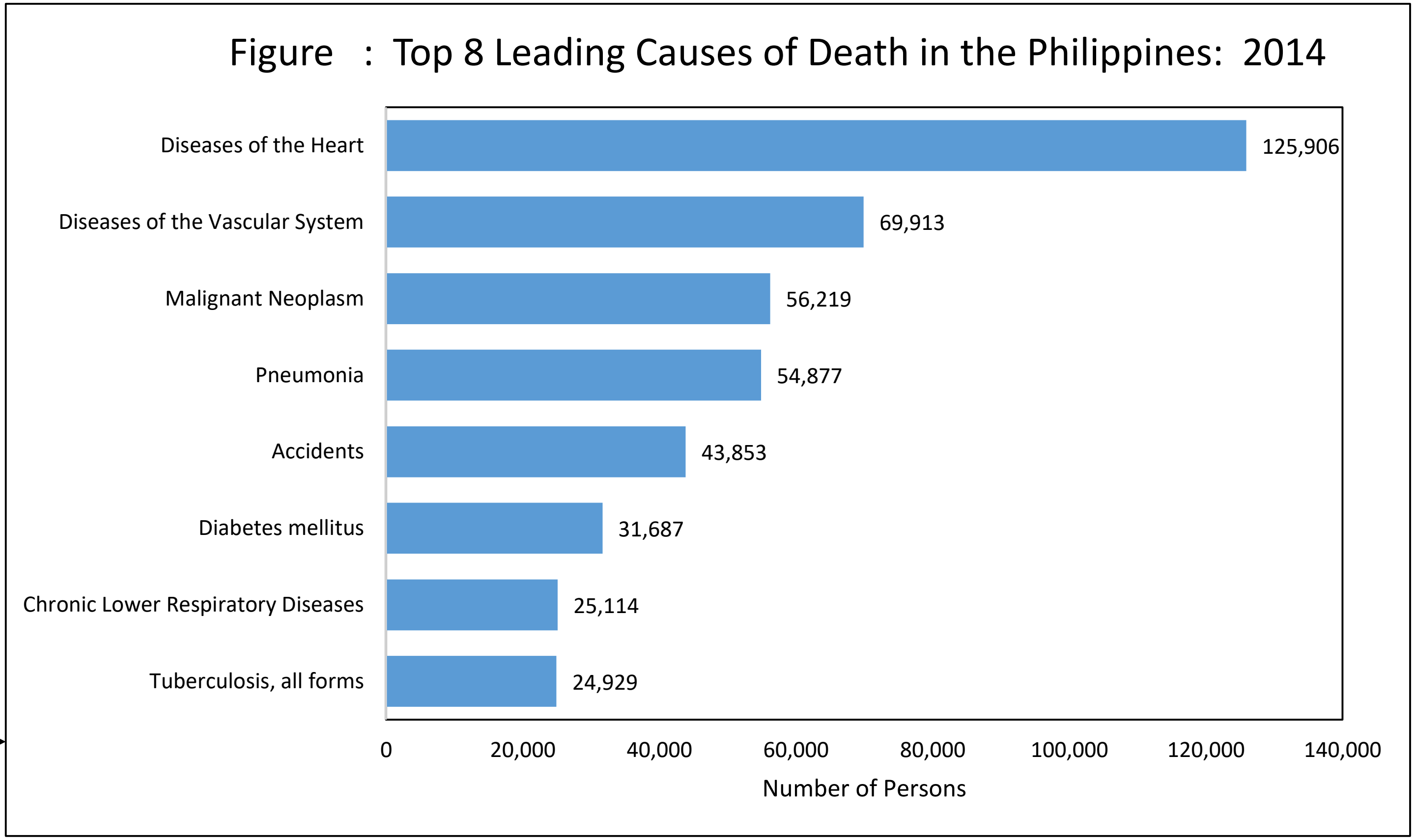
# Methods of Data Presentation

**Chart title**

**Axis Title**

**Scale figure**

**Source note**



Source: Department of Health

**Figure 10. Different Parts of a Simple Horizontal Bar Chart**

# Misconception on Graphical Presentation

- **Misconception:** The pie chart can be used for any number of categories.

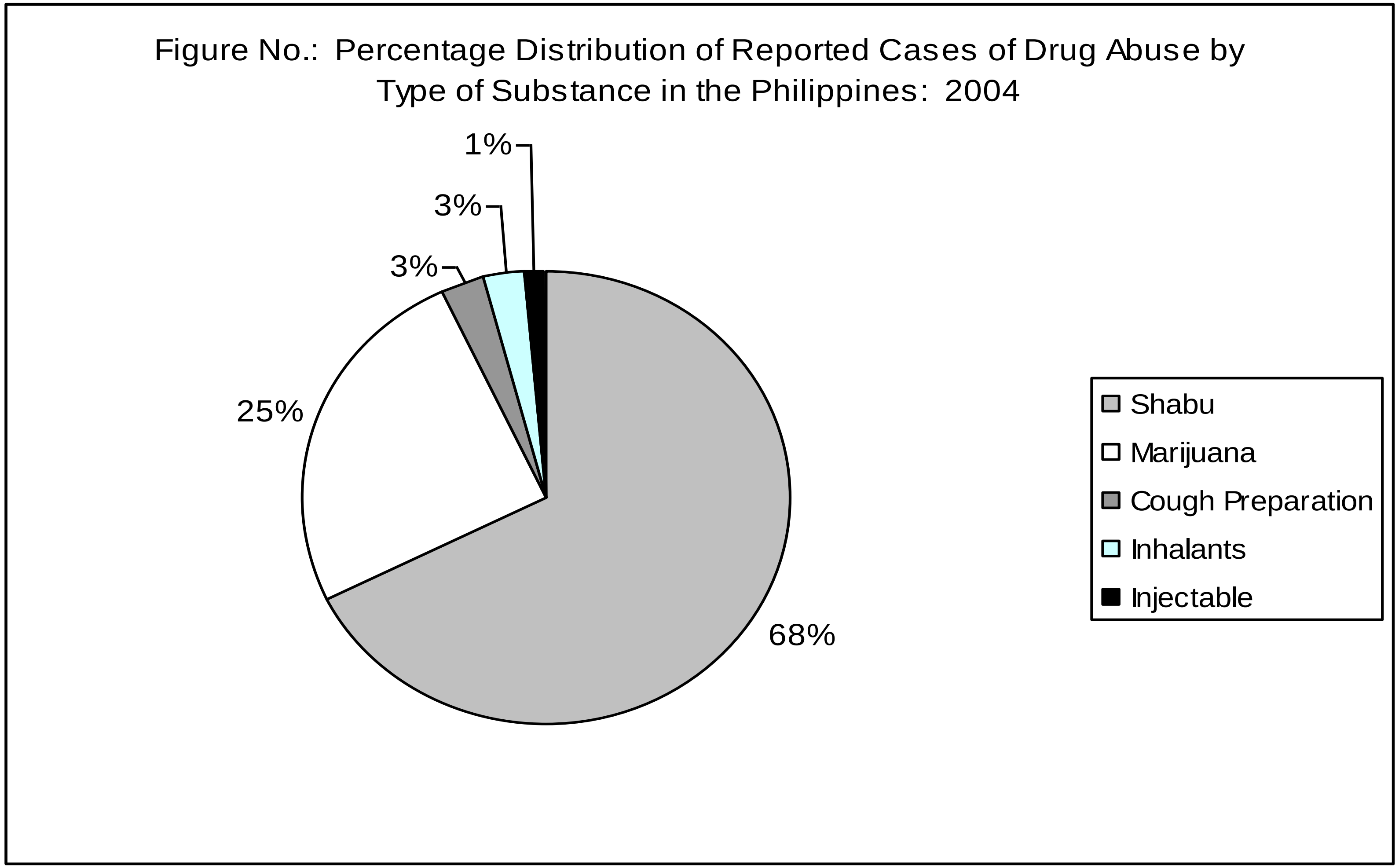
- The pie chart should be used for 8 categories or less.

# Misconception on Graphical Presentation

- **Misconception:** There is no arrangement for the different pie slices.
- The biggest pie slice should begin at 12oclock and the rest of the pie slices are arranged according to magnitude.
- The Others category should be placed last.

# Methods of Data Presentation

## Figure 1. Illustration of Pie Chart



# Misconception on Graphical Presentation

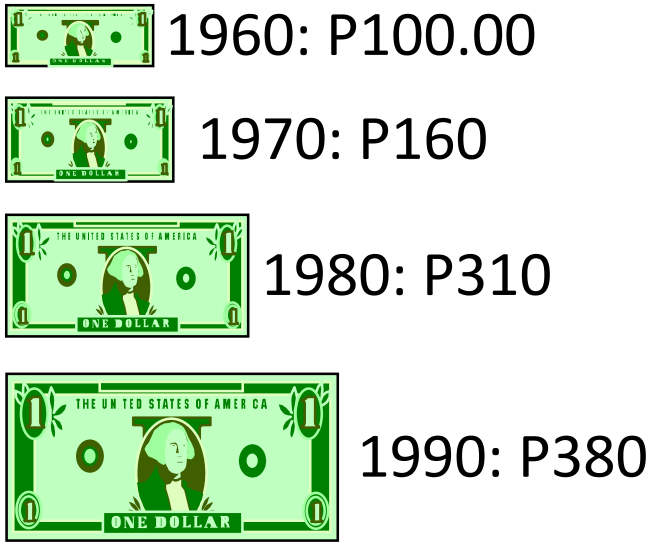
- **Misconception:** As the number becomes bigger, the symbol or picture used in a pictograph also becomes bigger.
- The symbol or picture used in a pictograph is constant in size. One symbol or picture is equivalent to a number.

# Methods of Data Presentation

## “Chart Junk”

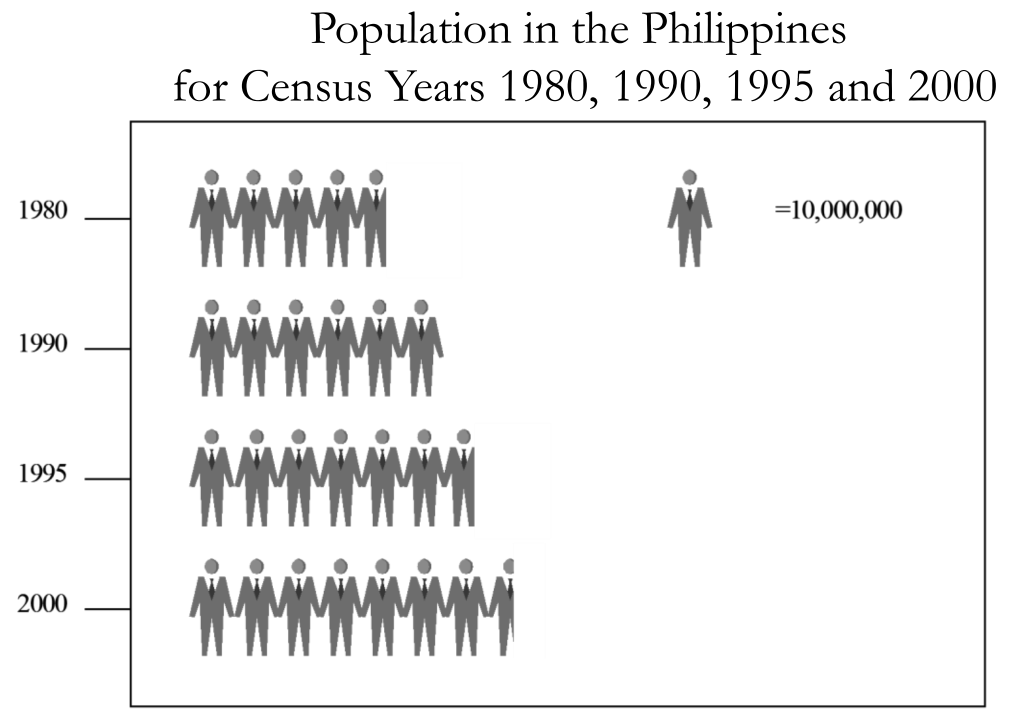
Bad Presentation

Minimum Wage



# Methods of Data Presentation

## Figure 13. Illustration of Pictograph



Source: Philippine Statistics Office

# Methods of Data Presentation

## Line Chart

- appropriate for time series data
- emphasis is on the movement
- shows trends, patterns, forecasts
- applicable for one or more time series data for comparison purposes

# Methods of Data Presentation

## Column Charts

- appropriate for comparing the magnitudes of variable in the x-axis for the different categories of variable in the y-axis
- for time series data, emphasis is on the magnitude and not the movement or trend
- The usual space between bars is around one-fourth of the width of the column.

# Methods of Data Presentation

## Horizontal Bar Charts

- for qualitative types of data given a specific time
- to compare the magnitudes of the different categories of a qualitative variable
- place the categories of the qualitative variable on the y-axis and the amount or number is on the the horizontal axis
- the spaces in between the bars may be one-fifth to one-half the width of the bar

# Methods of Data Presentation

## Pie Chart

- useful for data sorted into categories for a specific period
- emphasis is to show the components parts with respect to the total in terms of the percentage distribution
- use the pie chart if there are less than 8 categories in the data set

# Methods of Data Presentation

## Guidelines on Pie Chart:

- plot the biggest slice at 12 o clock
- arrange components of the pie chart according to magnitude
- if there is an “Others” category, put it in the last section
- use different colors, shadings, or patterns to distinguish one section of the pie to the other sections

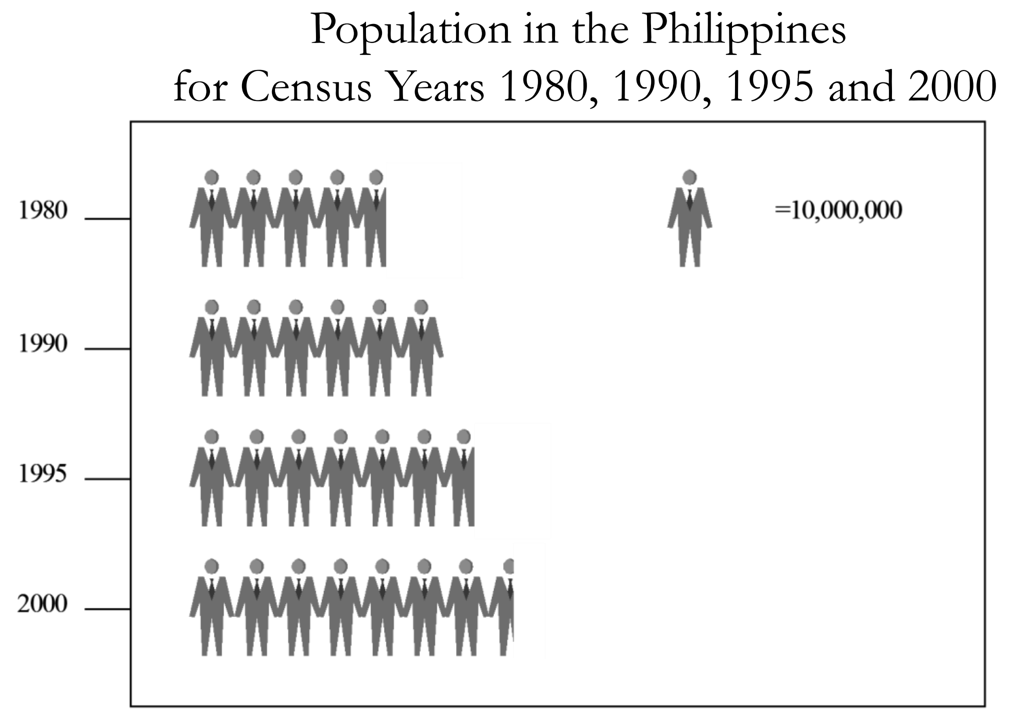
# Methods of Data Presentation

## Pictograph

- Appropriate for data sorted into categories for a specific period of time
- Gives an approximation only of the actual figures
- Compares the different categories
- Symbols selected should be self-explanatory and easy to understand
- Each symbol represents a number

# Methods of Data Presentation

## Figure 13. Illustration of Pictograph



Source: Philippine Statistics Office

# Misconception on Graphical Presentation

- **Misconception:** Variability of the data set is focused on the varying heights of the bars of the histogram.
- Focus should not be in frequencies but on data values.

# Misconception on Graphical Presentation

- **Misconception:** The median is interpreted to be the middle of the horizontal axis on a histogram.

- The histogram is the graphical presentation of a grouped data set. To get the median, estimate it by using the frequencies and the class boundaries.

# Misconception on Graphical Presentation

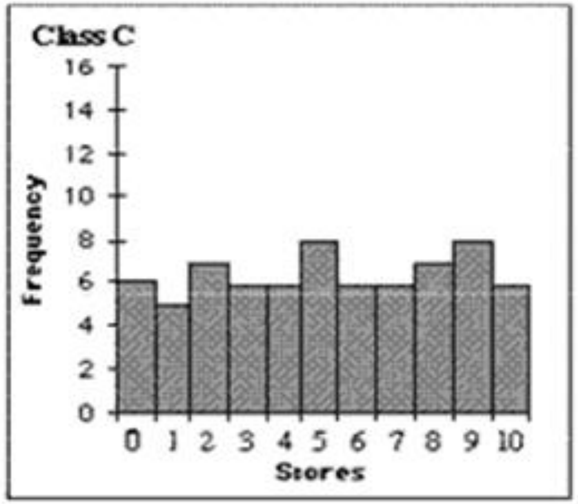
- **Misconception:** Vertical bar charts are interpreted as histograms wherein the shape is being described as skewed or symmetric.

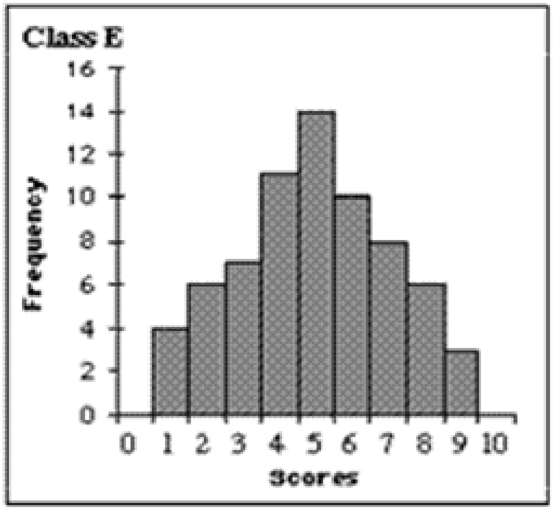
- Histograms are applicable only for quantitative data. A histogram doesn’t show data over time — it shows all the data at one point in time.

# Misconception on Graphical Presentation

- **Misconception:** A flatter histogram equates to less variability in the data.

- A flat histogram means that the data are spread out across the spectrum, hence a high variability.
- A histogram with a big lump in the middle and tails sloping sharply down on each side actually has less variability than a histogram that's straight across.





# Misconceptions on Summary Measures

# Misconception on Summary Measures

- **Misconception:** If one large group is made up of two subgroups (e.g., males and females), and if the mean score for each subgroup is available on a variable of interest, then the mean for the full group can be computed as the mean of the two subgroup means.
- The number of observations per group should be considered. Combined mean formula should be used.

# Misconception on Summary Measures

- **Misconception:** The median is the middle value of a set of observations.

- The median is the middle value of a set of ordered observations.

# Misconception on Summary Measures

- **Misconception:** The formula of the median is given as:

Case 1: n is odd

$$Median = X_{\frac{n+1}{2}}$$

Case 2: n is even

$$Median = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

- The formula of the median should be:

$$Median = X_{\left(\frac{n+1}{2}\right)}$$

$$Median = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

# Interpretation of Median

## Examples:

1. The following are the total receipts of 7 mining companies (in million pesos)

   1.2, 4.5, 6.5, 7.2, 10.4, 12.5, 50.6

   The median is 7.2.

Fifty percent of the seven mining companies have total receipts lower than 7.2 million pesos.

# Interpretation of Median

## Examples:

2. The following are the total receipts of 7 mining companies (in million pesos)

1.2, 4.5, 7.2, 7.2, 7.2, 12.5, 50.6

The median is 7.2.

At least fifty percent of the seven mining companies have total receipts lower or equal to 7.2 million pesos.

# Misconception on Summary Measure

- **Misconception:** a standard deviation indicates the “standard” amount that individual numbers deviate from the group’s mean, disregarding whether the original scores are above or below the mean.

- the SD is about 1.25 times as large as the “average deviation from the mean.” The SD is larger because it gives greater weight to scores that lie farther away from the mean. It does this by squaring the deviations. The SD is computed as the “root-mean-squared-deviation”

# Misconception on Summary Measure

- **Misconception:** the variability of one data set can be said to be large or small depending upon the value of the standard deviation.

- If there is only one data set, we cannot say that the variability is large or small based on the standard deviation since its value follows the magnitude of the observations.

# Misconception on Summary Measure

- **Misconception:** the standard deviation is used to determine which of two or more data sets is more variable.

- If two or more data sets have the same means and the same units of measurement, use the standard deviation to determine which data set is more variable.
- If two or more data sets have different means or different units of measurement, use the coefficient of variation to determine which data set is more variable.

# Two Measures of Dispersion

1. Measures of Absolute Dispersion
   - carries the unit of measure of the observations

2. Measures of Relative Dispersion
   - unitless so it can be used to compare the dispersion of two or more data sets with different means or different units of measurement.

# Measures of Position

A **measure of position** indicates the relative position of an observation in the distribution.

- Percentiles
- Quartiles
- Deciles

# Measures of Position

## Example of Percentiles:

- The 80th percentile of a distribution is a value such that at least 80 percent of the ordered observations are less than its value and at least 20 percent of the ordered observations are larger than its value.
- If $P_{80} = 75$: At least 80% of the ordered observations are less than 75.

  **OR** At least 20% of the ordered observations are larger than 75.

# Measures of Position

## Example of Percentiles:

- So any observation that is smaller than $P_{80}$ value belongs in the lower 80% of the distribution while any observation greater than $P_{80}$ value belongs in the upper 20% of the distribution.

# Measures of Position

## Quartiles

- Quartiles divide the ordered observations into 4 equal parts.

- 1st Quartile = 25th percentile
- 2nd Quartile = 50th percentile
- 3rd Quartile = 75th percentile

# Measures of Position

## Deciles

- Divide the ordered observations into 10 equal parts.
- Each part contains 10 percent of the observations.
- There are nine deciles and these are $D_1$, $D_2$, $D_3$, . . . , $D_9$.
- $D_1 = P_{10}, D_2 = P_{20}, \ldots, D_9 = P_{90}$

# Income Deciles

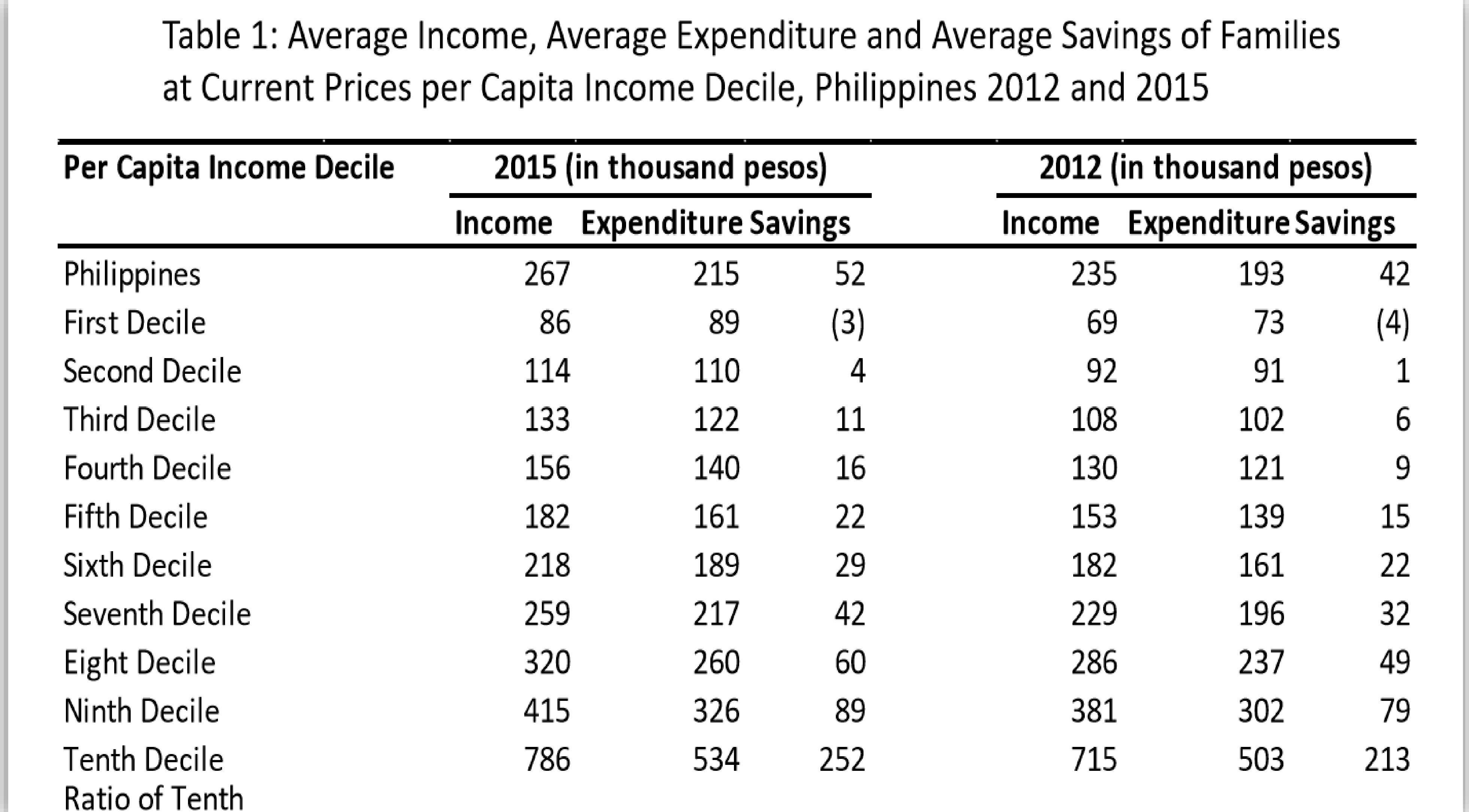


Table 1: Average Income, Average Expenditure and Average Savings of Families at Current Prices per Capita Income Decile, Philippines 2012 and 2015

| Per Capita Income Decile | 2015 (in thousand pesos) | | | 2012 (in thousand pesos) | | |
|---|---|---|---|---|---|---|
| | Income | Expenditure | Savings | Income | Expenditure | Savings |
| Philippines | 267 | 215 | 52 | 235 | 193 | 42 |
| First Decile | 86 | 89 | (3) | 69 | 73 | (4) |
| Second Decile | 114 | 110 | 4 | 92 | 91 | 1 |
| Third Decile | 133 | 122 | 11 | 108 | 102 | 6 |
| Fourth Decile | 156 | 140 | 16 | 130 | 121 | 9 |
| Fifth Decile | 182 | 161 | 22 | 153 | 139 | 15 |
| Sixth Decile | 218 | 189 | 29 | 182 | 161 | 22 |
| Seventh Decile | 259 | 217 | 42 | 229 | 196 | 32 |
| Eight Decile | 320 | 260 | 60 | 286 | 237 | 49 |
| Ninth Decile | 415 | 326 | 89 | 381 | 302 | 79 |
| Tenth Decile | 786 | 534 | 252 | 715 | 503 | 213 |
| Ratio of Tenth | | | | | | |

# Measures of Skewness

Interpretation:

SK = 0: symmetric $\bar{X} = \text{Md} = \text{Mo}$

SK > 0: positively skewed $\bar{X} > \text{Md} > \text{Mo}$

SK < 0: negatively skewed $\bar{X} < \text{Md} < \text{Mo}$

# Misconceptions on Expectations and Variance

# Misconceptions on Expectation and Variance

Evaluate Var(2X)

As E(2X)=2E(X)

Var(2X)= 2Var(X)

Confusion: Expectation and Variance

# Misconceptions on Expectation and Variance

Confusion: Expectation and Variance

Relationship:

$$\begin{aligned} Var(X) &= E[X-E(X)]^2 \\ &= E[X-\mu]^2 \end{aligned}$$

Proof:

$$\begin{aligned} Var(2X) &= E[2X-E(2X)]^2 \\ &= 4E[X-\mu]^2 \\ &= 4Var(X) \end{aligned}$$

# Misconceptions on Expectation and Variance

$$
\begin{aligned}
\mathrm{Var}(2X) &= E[2X-E(2X)]^2 \\
&= E[2X-2\mu]^2 \\
&= E[2X-2\mu]^2 \\
&= E[4X^2+4\mu^2-8X\mu] \\
&= 4E[X^2+\mu^2-2X\mu] \\
&= 4E[X-\mu]^2 \\
&= 4\mathrm{Var}(X)
\end{aligned}
$$

# Misconceptions on Standard Normal

# Misconceptions on Standard Normal

Is standardized data normally distributed ?

Standardization is used together with Normal distribution most of the time

# Standardization and Normal Distribution

- The shape of the z-score distribution will be exactly the same as the distribution from which it arose.
- That is, if the **original** distribution is positively **skewed**, the z-score distribution will also be positively **skewed**.
- However, all z-score distributions will have a **mean** of 0 and a standard deviation of 1.

# Misconceptions on Inferential Statistics

# Misconceptions on Confidence Interval

- **Misconception:** "There is a 95% chance that the true population mean falls within the confidence interval."
- **Misconception:** "The mean will fall within the confidence interval 95% of the time."
- **Misconception:** "The probability of the parameter falling somewhere between the end points of that interval is equal to .95".

# Misconceptions on Confidence Interval

**Correct Interpretation:**

- A 95% level of confidence means that 95% of the confidence intervals calculated from these random samples will contain the true population mean.

- In other words, if you conducted your study 100 times you would produce 100 different confidence intervals. We would expect that 95 out of those 100 confidence intervals will contain the true population mean.

# Misconceptions on Confidence Interval

## Notes on Confidence Interval:

- A confidence interval is not a probability, and therefore it is not technically correct to say the probability is 95% that a given 95% confidence interval will contain the true value of the parameter being estimated.

- Since the event (construction of the confidence interval) has already occurred, the true value of the targeted parameter either does or does not lie in the interval.

# Example of Correct Interpretation of Confidence Interval

**95% interval estimate for P: 0.828 < p < 0.872.**

- Correct: "We are 95% confident that the interval from 0.828 to 0.872 actually does contain the true value of the population proportion P.“

This means that if we were to select many different samples of size 100 and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population portion P.

# Misconceptions on Hypothesis Testing

- **Misconception:** The research hypothesis is a claim or statement on the statistic.

- The research hypothesis is a claim or statement on the parameter.

# Misconceptions on Hypothesis Testing

- **Example of Misconception:**

- Ho: $\bar{X} = 45 \ \ vs$ Ha: $\bar{X} < 45$

**Correct Notation:**

- Ho: $\mu = 45 \ \ vs$ Ha: $\mu < 45$

# Misconceptions on Hypothesis Testing

- **Misconception:** We accept the null hypothesis if the alternative hypothesis cannot be supported.

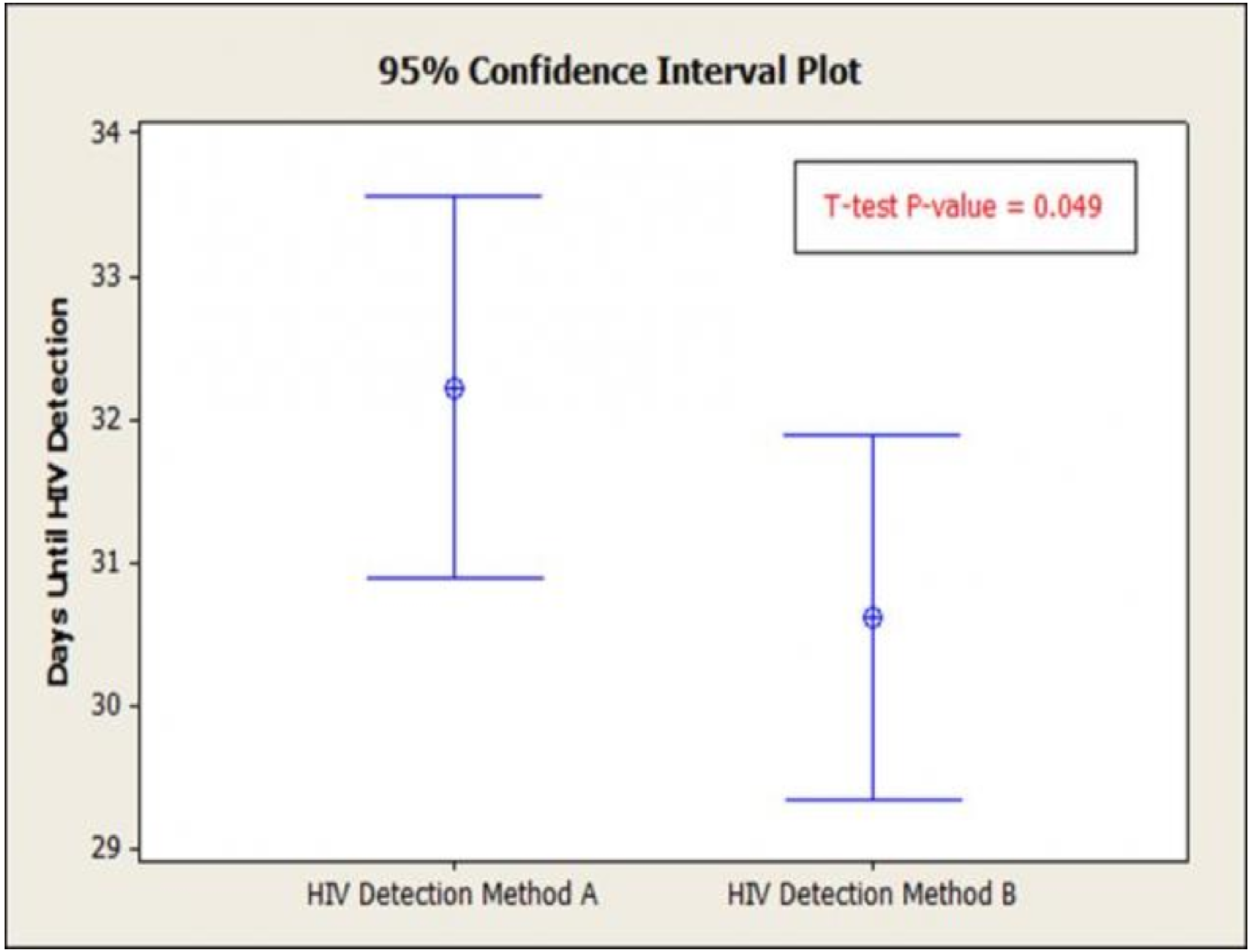
- We do not reject the null hypothesis since there is no sufficient sample evidence to reject the null hypothesis.
- OR We do not reject the null hypothesis. There is no sufficient sample evidence to support the alternative hypothesis.

# Misconceptions on Hypothesis Testing

- **Misconception:** When 95% confidence intervals for the means of two independent populations overlap, there is no significant difference between the means.

- When 95% confidence intervals for the means of two independent populations do not overlap, there will indeed be a statistically significant difference between the means (at the 0.05 level of significance).
- However, the opposite is not necessarily true. CI's may overlap, yet there may be a statistically significant difference between the means.

# Mistake #1: Misinterpreting Overlapping Confidence Interval



Two 95% confidence intervals that overlap may be significantly different at the 95% confidence level. **What's the significance of the t-test P-value?** The P-value in this case is less than 0.05 (0.049 < 0.05), telling us that there is a statistical difference between the means, (yet the CIs overlap considerably).

# Misconceptions on Hypothesis Testing

- **Misconception:** "When we make a decision in hypothesis testing, we have proven which hypothesis is true“
- Hypothesis testing is **not** a formal proof of any kind and at no stage proves that any hypothesis is true.
- It only provides evidence to *support* the decision whether or not to reject the assumption that the Null Hypothesis is true.

# Misconceptions on Hypothesis Testing

- **Misconception:** "If we reject the Null Hypothesis then the Alternative Hypothesis must be true“

- Reject Ho at $\alpha = 0.05$. There is sufficient sample evidence to support the alternative hypothesis.
- OR Reject Ho $\alpha = 0.05$ . There is sufficient sample evidence to reject the null hypothesis.

# Misconceptions on Hypothesis Testing

- **Misconception:** "A statistically significant result is important“
- The word significant, when used in statistics, simply means that there was sufficient sample evidence to reject the assumption that the Null Hypothesis is true, it does not say anything about the importance of the result.

# Misconceptions on Hypothesis Testing

- **Misconception:** "We should only publish statistically significant results“

- Sometimes referred to as the "publication bias" this misconception is actually contrary to good scientific practice.
- ***ALL*** results should be published, even those of experiments that didn’t work.

# Misconceptions on Hypothesis Testing

- **Misconception:** "Confidence intervals are better than hypothesis tests"

- Confidence interval is just a different form of decision statistic you can use in a hypothesis test.
- It gives us an interval estimate for the population parameter based upon the sample data.
- If the interval captures the null hypothesized value then the result is not statistically significant.

# Misconceptions on Hypothesis Testing

- **Misconception:** "Smaller *p*-values indicate a bigger effect"

- A *p*-value **does not** measure anything about the effect or size of the difference observed.
- The *p*-value simply measures the probability of observing the sample data under the assumption that the Null Hypothesis is true.
- Many things can contribute to getting a small or large *p*-value, including sample size and/or measurement precision.

# Misconceptions on Hypothesis Testing

- **Misconception:** "A *p*-value near the significance level can be interpreted as *approaching* a significant or non-significant result"

- The *p*-value cannot be interpreted as *approaching* anything.
- The *p*-value is simply the probability of observing the sample data under the assumption that the Null Hypothesis is true.
- If the *p*-value is near the significance level then you may want to re-run your study with a different sample if you want to confirm whether or not the decision you made is repeated — this is what is meant by replication.

# Misconceptions on Hypothesis Testing

- **Misconception:** A low probability value indicates a large effect.

  Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true.
- A low probability value can occur with small effect sizes, particularly if the sample size is large.

# Misconceptions on Hypothesis Testing

- **Misconception:** A non-significant outcome means that the null hypothesis is probably true.

  Proper interpretation: A non-significant outcome means that there is no sufficient sample evidence to reject the null hypothesis.

- **Thank you very much for listening.**

**PHILIPPINE STATISTICAL RESEARCH AND TRAINING INSTITUTE**
7TH FLOOR, SOUTH INSULA CONDOMINIUM
61 TIMOG AVENUE, BARANGAY SOUTH TRIANGLE, DILIMAN, QUEZON CITY

CONTACT NUMBERS: 288.4948/426.0620/929.7543/288.4150/374.4587/920.9649

WWW.PSRTI.GOV.PH