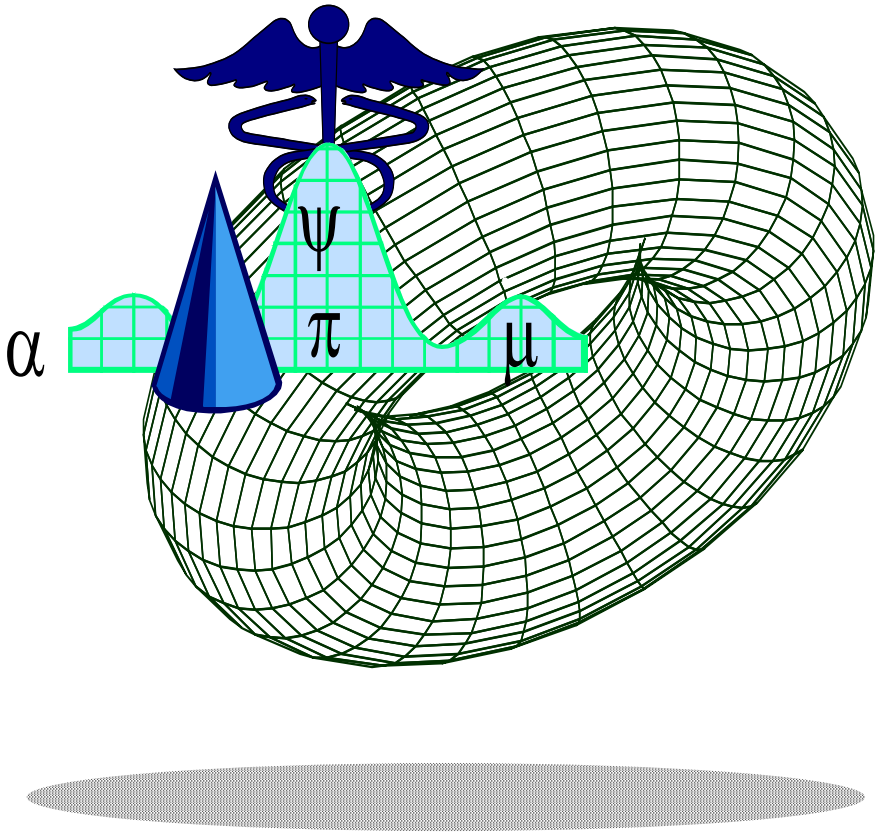
# WORKBOOK FOR BIOSTATISTICS

## 24 - 28 September 2012




**Name of participant:……..**

# Workbook for Biostatistics

**Topics covered:**
Overview of statistics

*Prepared by:*
Bandit Thinkhamrop, Ph.D. (Statistics)
Associate Professor of Biostatistics
Department of Biostatistics and Demography, Faculty of Public Health
Khon Kaen University, THAILAND

E-mail: bandit@kku.ac.th

# INTRODUCTION:

A good research needs to have involved an important issue, had useful objectives, used sensible design, used adequate sample size, performed appropriate analysis, and drawn reasonable inferences from the findings. Statistics plays a very important role not only on analysis but also on the other components.

> *Statistics is a curious amalgam of mathematics, logic and judgement. It is logical process that cause more difficult than mathematics- the principles of good design, and the concepts underlying data analysis and interpretation* (Altman , 1991; page 9).

This workbook is designed for participants who have a limited background in mathematics and statistics. It is believed that if participants can be convinced that statistics is a very important tool and useful for their professional, they will then open themselves for this difficult subject. Once they understood the key concept behind basic statistics, they will find their own way in trying to understand a more complicated statistical methods. An example of the evidence of success, reported by a participant of a previous a course in which this workbook had been used, is as quoted - "*The class has been finished long time ago but I found many of my classmates keep on reading statistics books. I bought several books of this kind for myself in which I have never even taken a glance on it prior to the course."* However, participants who gained a little from this is not abnormal.

# OBJECTIVES:

Upon completion of this workbook, participants should be able to:

1. Describe the underlying "core" concepts of statistical methods.
2. Outline the principle of choosing appropriate statistical methods
3. Outline important steps in data analysis including examining the data, describing the study sample, and answering the research questions
4. Describe the "core" concepts underlying statistical inference
5. Interpret correctly the confidence intervals and the p-value.

# CONTENTS:

**PART 1**: Pre-test ..... 3
**PART 2**: Overview of statistics ..... 4
**PART 3:** Examining distribution of the data ..... 14
**PART 4:** Descriptive statistics for describing the study sample ..... 17
**PART 5:** Statistical inference for answering the research question ..... 19
**PART 6:** Confidence interval vs. p-value ..... 26
**PART 7**: Comprehensive exercise ..... 27
**PART 8**: Post-test ..... 31

# **PART 1**: PRETEST

**A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo (p-value < 0.05).**

*Which of the following statement do you prefer?*

4[ ]1. It has been proved that treatment is better than placebo.

2[ ]2. If the treatment is not effective, there is less than a 5 percent chance of obtaining such results.

6[ ]3. The observed effect of the treatment is so large that there is less than a 5 percent chance that the treatment is no better than placebo.

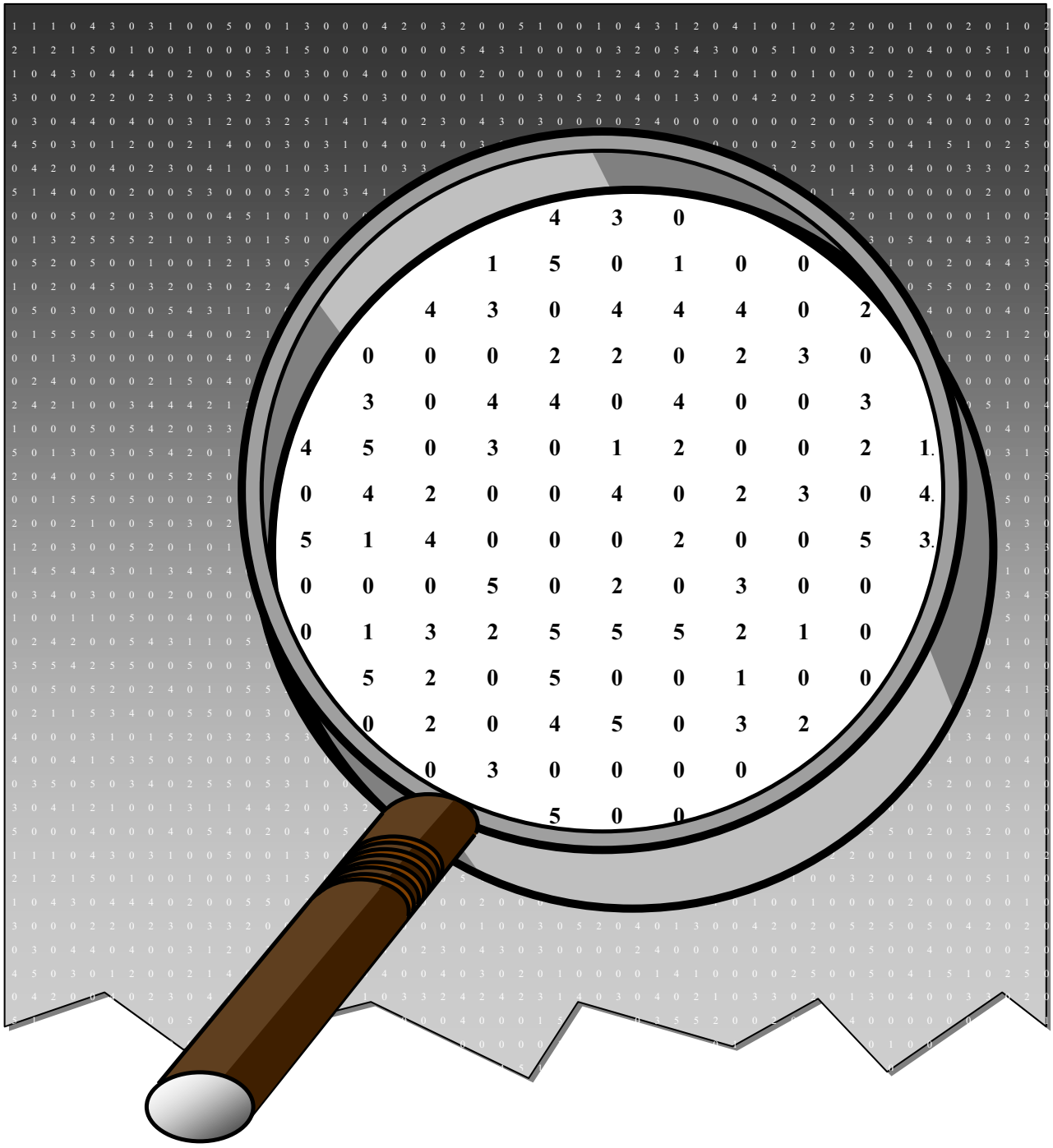
0 [ ]4. I do not really know what a is and do not want to guess.

***Source:*** *Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987) What do doctors know about statistics?. Statistics in Medicine, **6**, 3-10.*

Justifications of the answer:

# PART 2: OVERVIEW OF STATISTICS

1. Followings are parts of unlimited item of data. Let we call these the ***target population***. The data represent the number of cigarettes smoked per day reported by each member of a community. Assuming these data are located at random in this long piece of paper, lists of items within the magnifying glass are what we can clearly see, i.e. what we only have at hand. Let we call these the ***sampling frame***. Now please randomly selected your study sample from the sampling frame with the sample size of 10.



2. List all of your samples (n = 10)

3. Cast the data in the table format

| Subject No. | Number of cigarettes per day |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

4. If we had also recorded age and sex of the subjects during data collection, the table would looked like what is shown below. We now replace the heading with some shorter word to facilitate further steps in handling the data.

These headings are called …….. ([ ]Data / [ ]Variable).

The numbers filled under each heading are called …….. ([ ]Data / [ ]Variable).

| ID | CIGAR | AGE | SEX |
|---|---|---|---|
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| 9 | | | |
| 10 | | | |

This is the data displaying format similar to what we always see in the computer.

5. We need to have description of the data for further data processing and analysis. It is known as ***data dictionary*** which allows anybody to understand what the data is about. It is necessary and researchers are recommended to make it available.

| Variable name | Descriptions | Values |
|---|---|---|
| ID | Identification number | 1 to n |
| CIGAR | Cigarettes smoked per day | Number of cigarettes |
| AGE | Age | Age in years |
| SEX | Sex | 1 = Male; 2 = Female |

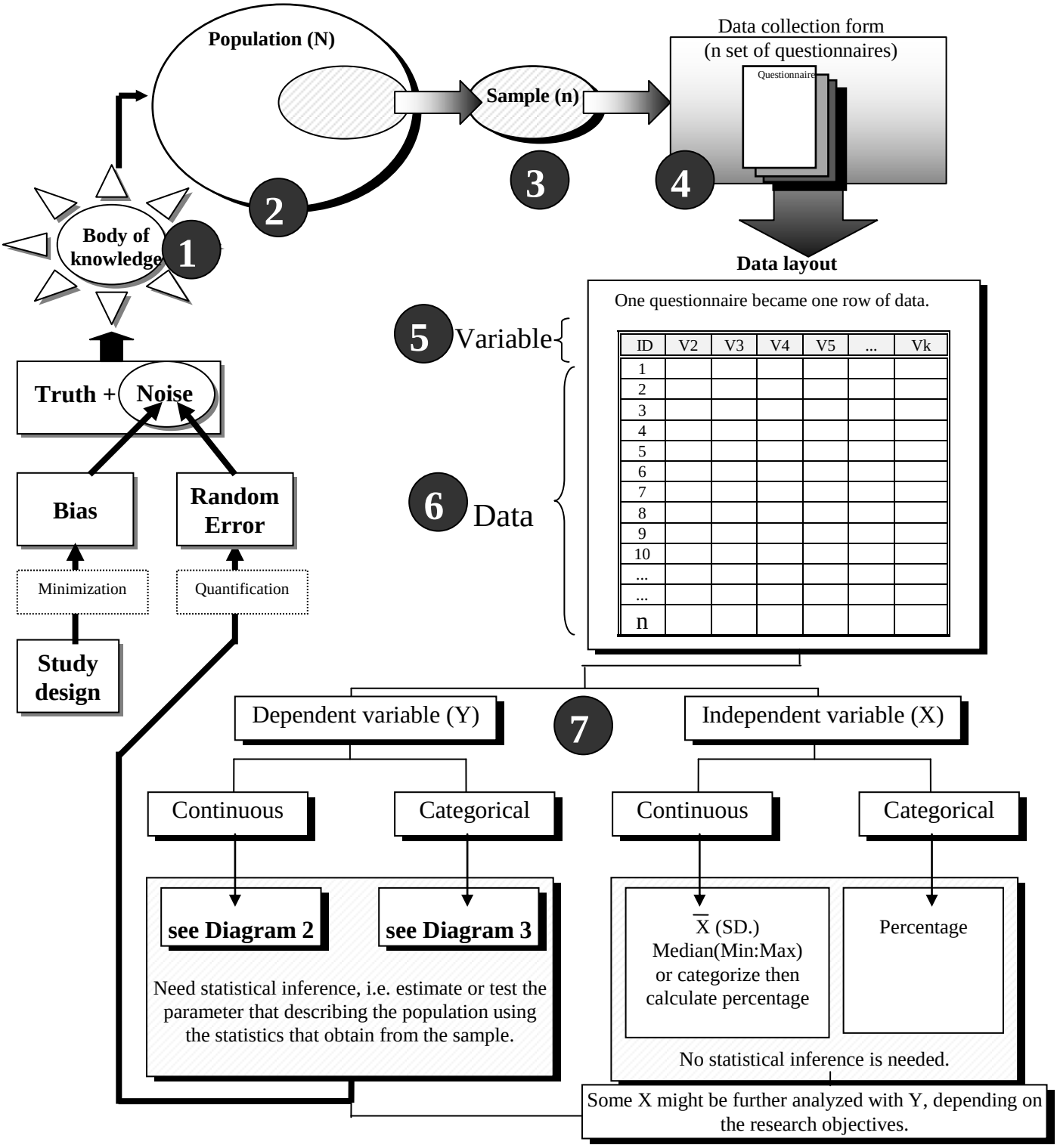
**Example of some data:**

| ID | CIGAR | AGE | SEX |
|---|---|---|---|
| 1 | 0 | 40 | 1 |
| 2 | 5 | 23 | 1 |
| 3 | 0 | 35 | 2 |
| ---------- n - 4 records omitted ---------- | | | |
| n | 0 | 42 | 2 |

6. Outline how the questionnaire would look like.

7. All previous steps can be summarized by parts of Diagram 1 shown below.

# Diagram 1
## OVERVIEW OF STATISTICS



8. Based on Diagram 1, there are some interesting topics to address.

8.1 **1** is the body of knowledge which is the goal of all research. We can take the Pre-test as an example. What you have been trying to answer the pre-test question is really that you are trying to draw a conclusion from the study. It's the knowledge gain from reading the paper.

**A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo (p-value < 0.05).**

*Which of the following statement do you prefer?*

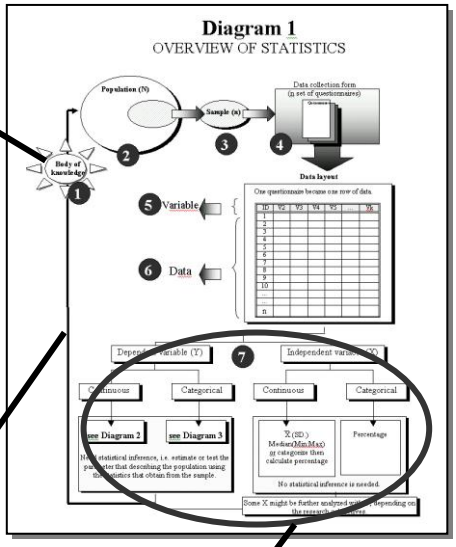
[ ]1. It has been proved that treatment is better than placebo.

[ ]2. If the treatment is not effective, there is less than a 5 percent chance of obtaining such results.

[ ]3. The observed effect of the treatment is so large that there is less than a 5 percent chance that the treatment is no better than placebo.

[ ]4. I do not really know what a p-value is and do not want to guess.

***Source:*** *Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987) What do doctors know about statistics?, Statistics in Medicine, 6, 3-10.*




**Statistical inference:**
Describe characteristics of the population
*(i) Estimation*
*Commonly reported as 95% Confidence intervals*
*(ii) Hypothesis testing*
*Commonly reported as p-value*
*(or significant / not significant)*

**Descriptive statistics:**
Describe characteristics of the sample.

**Question:**

8.1.1 Regarding the Pre-test question,
what have been reported? ……

what is lacking? ……...

8.1.2 If you are to decide whether or not the new treatment should be used, what additional evidence you need? ……..
……
……

8.1.3 In the context of the smoking survey, what should be the body of knowledge? ……..
……
……
……

8.2 2 and 3 are about the population to which the findings from the study will be applied.

**Question**: In the smoking survey, suppose that we aim to estimate the prevalence (percentage) of smoking in a community.

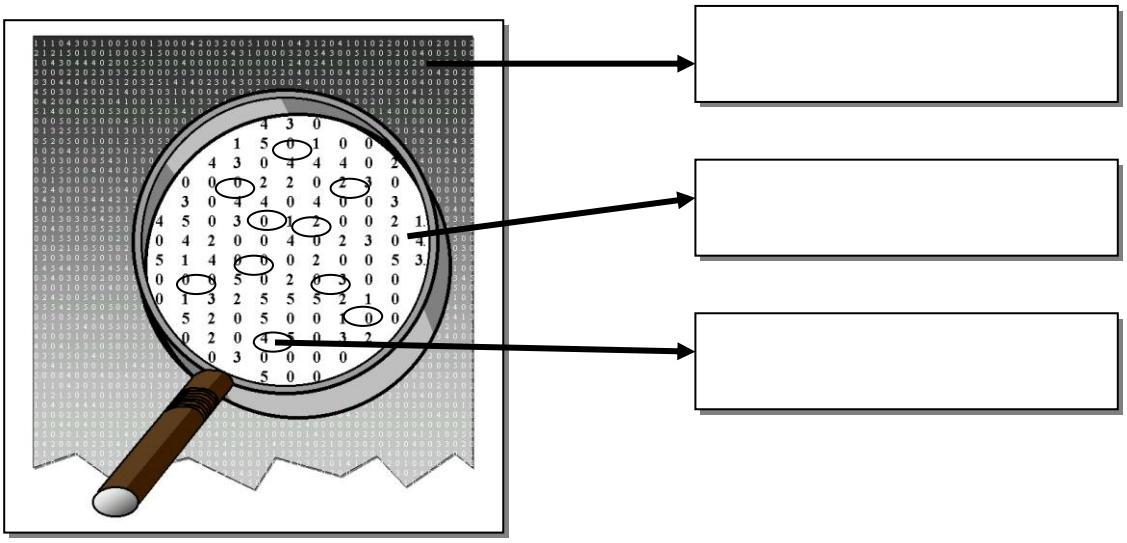
8.2.1 The target population is ……..……..

8.2.2 The sampling frame is ……..…….

8.2.3 The sample is……..……...

8.2.4 The sample size is ……..……..…….

8.2.5 The sampling method is ……..……..



8.3 4 to 7 are about data collection. Then data processing which (always) use the computer followed by data analysis.

Now put in the context of the smoking survey. If ask "how can we do with these data?", we need to first answer the question - "What is the objective of the study?"

Let suppose we have the data only CIGAR, AGE, and SEX that were in #5 shown above. There may be several possible type of research objectives. Followings are some commonly found, follows by lists of questions needed to answer so that one will know how to analyse the data.

8.3.1 The study is to ***estimate*** the average number of cigarettes.

Dependent variable: ……. ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the dependent variable:…… ([ ]Continuous / [ ]Categorical)

Independent variable(s): ……... ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the independent variable:……. ([ ]Continuous / [ ]Categorical)

The diagram to see further for identifying appropriate statistical methods is ……

8.3.2 The study is to quantify magnitude of relationship between smoking and age.

Dependent variable: ……. ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the dependent variable:…… ([ ]Continuous / [ ]Categorical)

Independent variable(s): ……... ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the independent variable:……. ([ ]Continuous / [ ]Categorical)

The diagram to see further for identifying appropriate statistical methods is ……

8.3.3 The study is to compare the prevalence of smoking between male and female.

Dependent variable: ……. ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the dependent variable:…… ([ ]Continuous / [ ]Categorical)

Independent variable(s): ……... ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the independent variable:……. ([ ]Continuous / [ ]Categorical)

The diagram to see further for identifying appropriate statistical methods is ……

8.3.4 The study is to determine factors affecting smoking.

Dependent variable: ……. ([ ]CIGAR / [ ]AGE / [ ]SEX)

Type of the dependent variable:…… ([ ]Continuous / [ ]Categorical)

Independent variable(s): ……... ([ ]CIGAR / [ ]AGE / [ ]SEX)

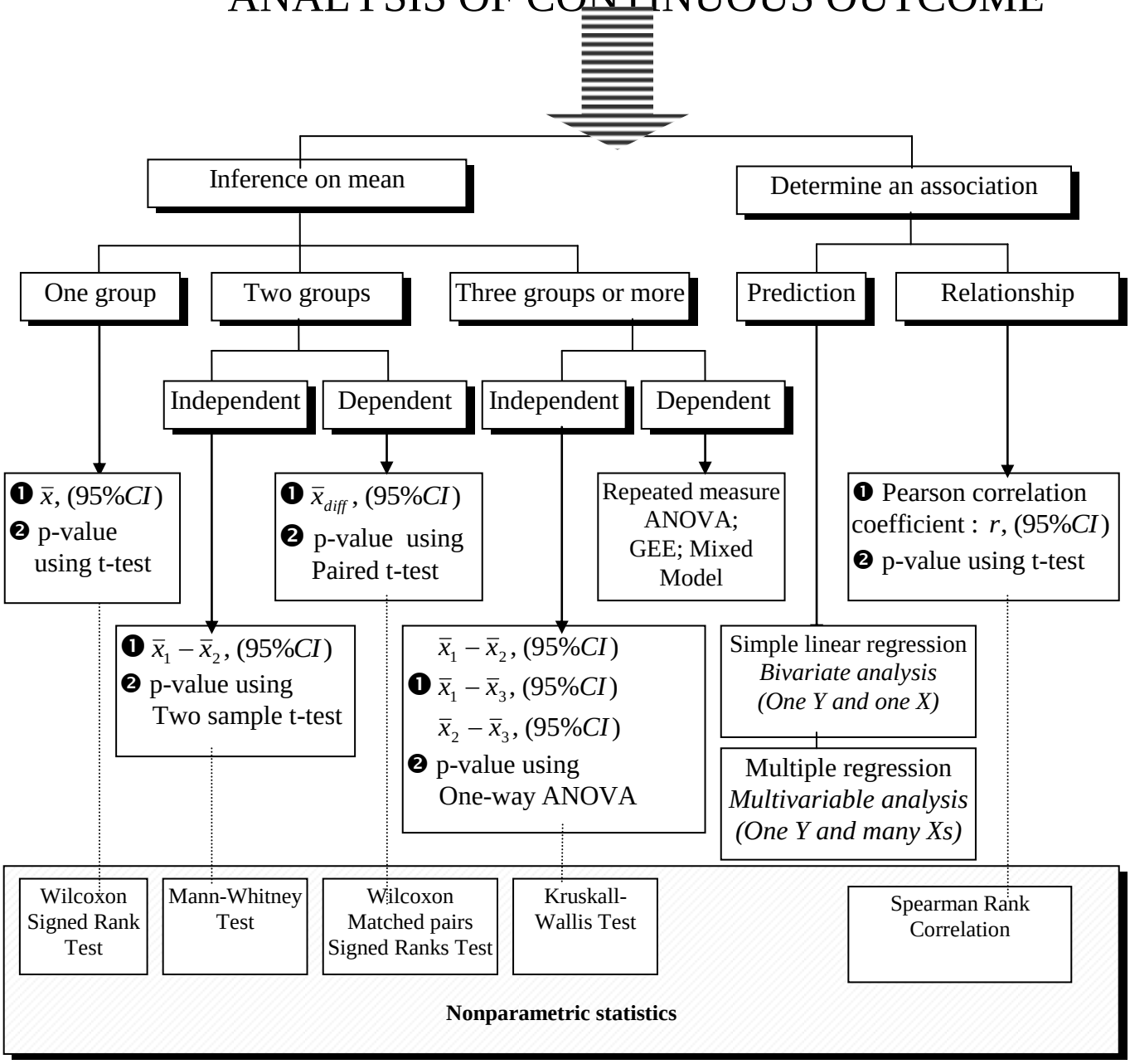
Type of the independent variable:……. ([ ]Continuous / [ ]Categorical)

The diagram to see further for identifying appropriate statistical methods is ……

9. Let's examine Diagrams 2 and 3 for identifying appropriate statistical methods.

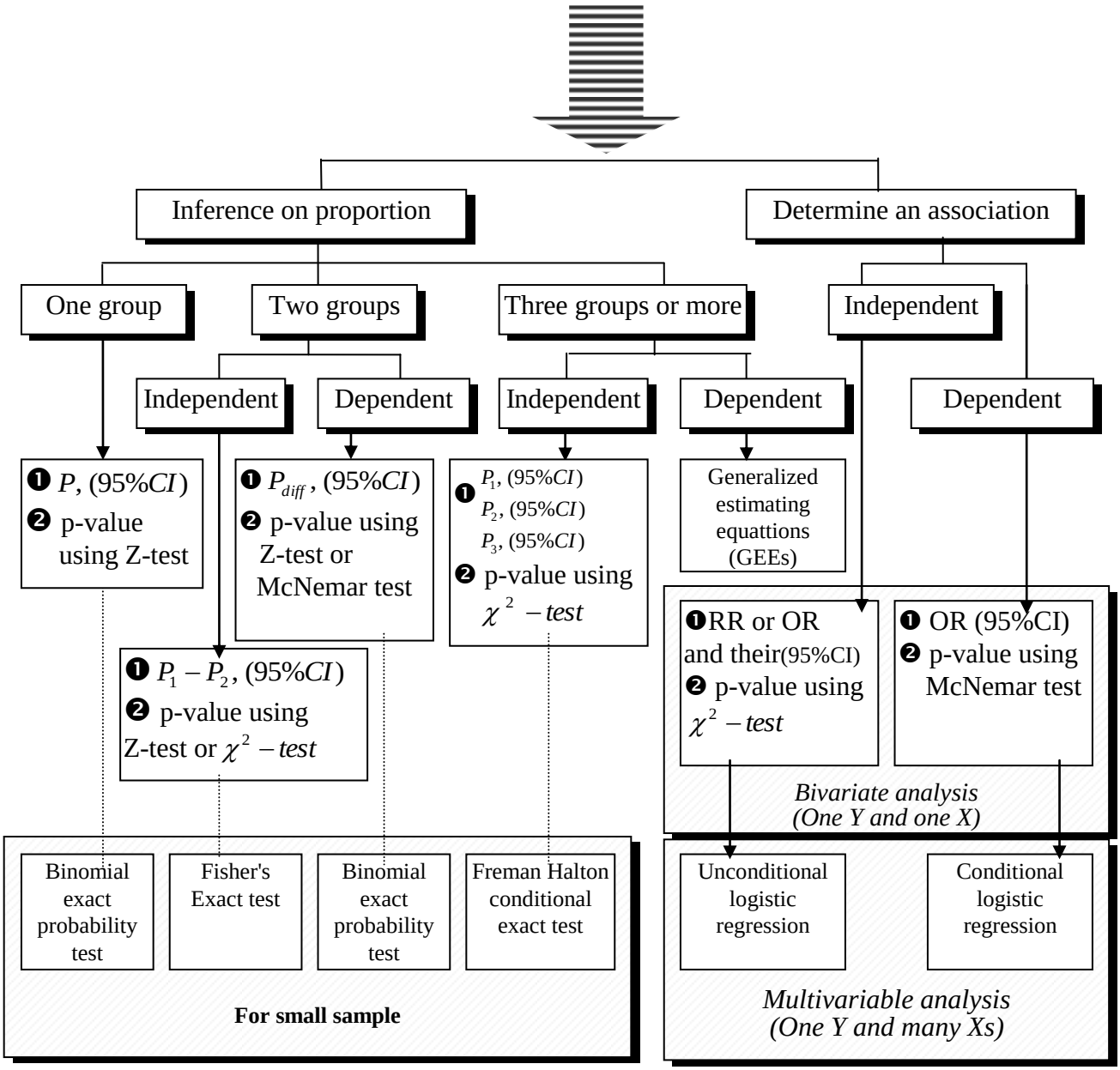
# Diagram 2

## ANALYSIS OF CONTINUOUS OUTCOME



# Diagram 3

## ANALYSIS OF CATEGORICAL OUTCOME



10. In reference to question 8, we have examined the diagram specified at the end of each question. Now, specify the statistical methods of your choice and your justifications.

10.3.1 The study is to ***estimate*** the average number of cigarettes.

Statistical methods: ……..

Justifications: ……

10.3.2 The study is to quantify magnitude of relationship between smoking and age.

Statistical methods: ……..

Justifications: ……

10.3.3 The study is to compare the prevalence of smoking between male and female.

Statistical methods: ……..

Justifications: ……

10.3.4 The study is to determine factors affecting smoking.

Statistical methods: ……..

Justifications: ……

10.3.5 Specify a research objective which requires paired t-test.

……..

# **PART 3:** EXAMINING DISTRIBUTION OF THE DATA

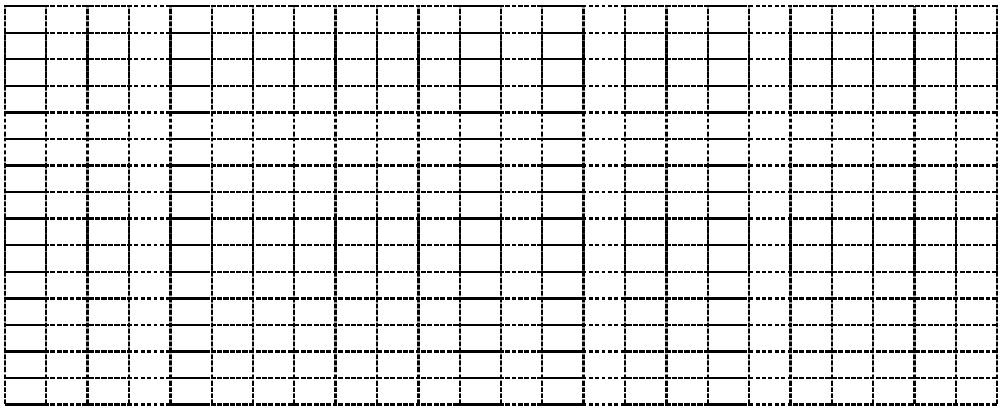
From the smoking survey you have conducted, we will pretend that the objective of is to estimate the average number of cigarette smoked per day. Since the dependent variable is continuous. Following steps is necessary.

1. Construct frequency table for weight of the student of n = 10

| **Number of cigarettes per day** | **Frequency** | **Percent** | **Cumulative percentage** |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| **Total** | | | |

**Question**: For how many percents of the sample smoke 3 or more cigarettes per day?…….%

2. Construct a histogram and then superimpose by a curve covering the same size of area, i.e., n blocks or 10 blocks for this example.



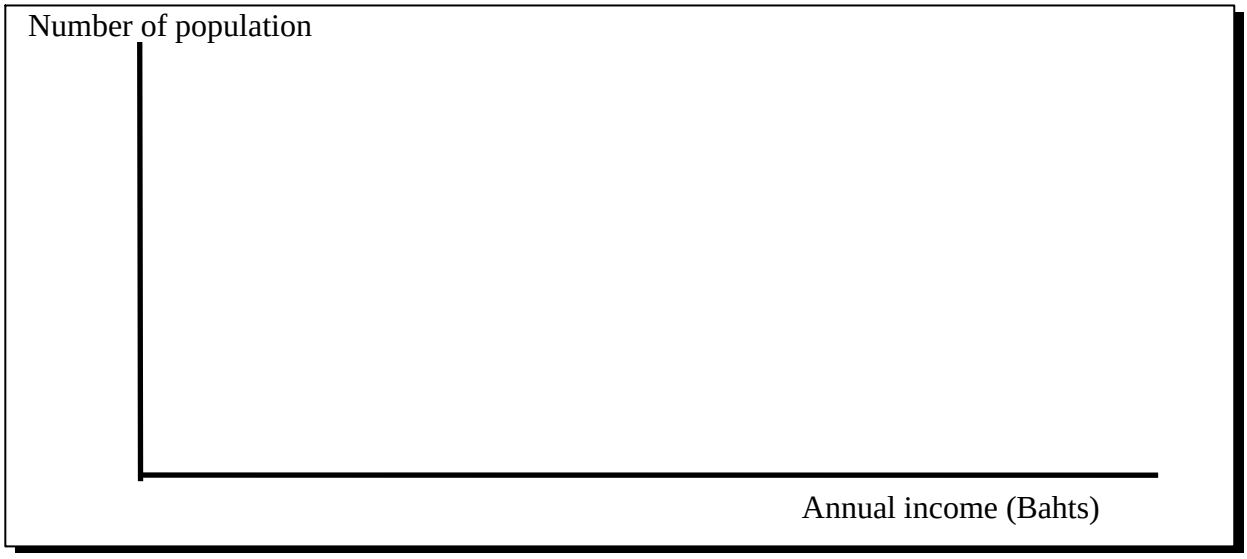
**Question**: For how many percents of the area under the curve corresponds to smoking of 3 or more extreme?…….%

If we can validly assume that the distribution in anywhere in Thailand is similar to what was found by this study, how many people smoking 3 or more cigarette per day within a community of 400,000 population? ……

3. Describe type of the distribution of weight
   [ ] Symmetry
   [ ] Left skew (negative skew)
   [ ] Right skew (positive skew)
   [ ] Others ……

4. Followings are hypothetical data ***"A study of annual income per capita among Thai people found that 80% were poor (poor is classified as those whose income of less than 20,000 bahts per year). The range is as small as 500 bahts to several thousand million bahts "***

4.1 Construct the approximate the frequency curve to reflect the findings




4.2 What are the income which make the area under the curve covering 80% of total Thai population.

.....

How did you get that:

.....

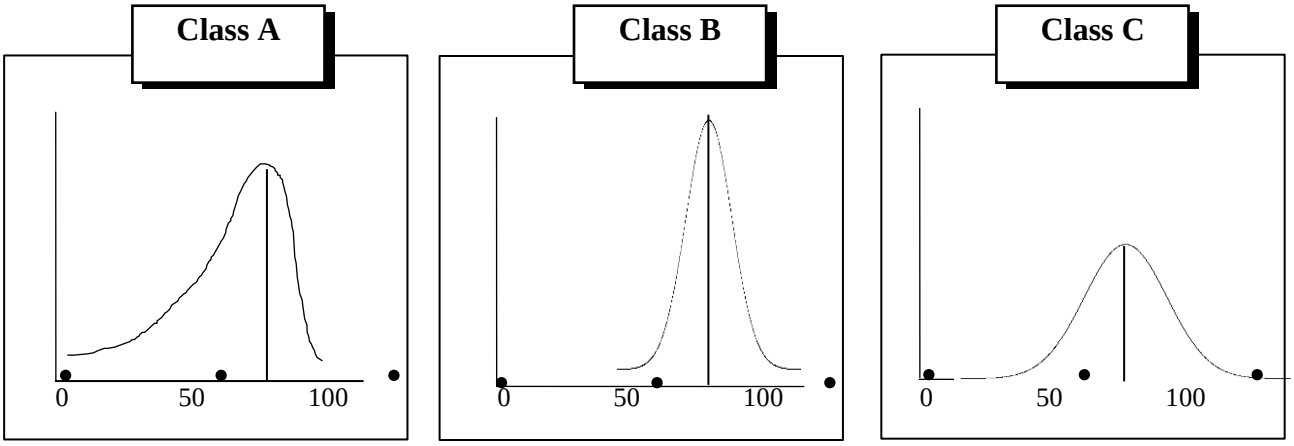
*(Note that this is similar to the question for parameter estimation, i.e. it is to obtain the lower and upper limit of a quantity that cover a specified magnitude of proportion or probability)*

4.2 What is the probability of having found people who have the annual income of 20,000 Bahts or higher: .....

How did you get that: .....

*(Note that this is similar to the question for testing hypothesis, i.e., it is to obtain the probability from a quantity which is calculated from comparing the observed estimate with the hypothesized value. Such quantity is the statistical test and the corresponding probability is the p-value which will then be compared with the cut point such as 0.05 and interpret as significant or non-significant.)*

5. Examine the frequency curve reflecting the marks of biostatistics examination for three classrooms bellow and then answer the questions:




5.1 Describe the types of the distribution and their implications

Class A .....

Class B .....

Class C .....

5.2 What is the main different of Class B and C

.....

5.3 What is the most appropriate summary statistics for each class

Class A .....

Class B .....

Class C .....

6. To summarize about a continuous data, the **two** important numbers to be reported are

6.1 …… 6.2 ……

7. Write on your own word regarding the important of examining the distribution of data

.....

.....

# **PART 4:** DESCRIPTIVE STATISTICS FOR DESCRIBING THE STUDY SAMPLE

In practice, this is the first part presented under the result section of the research report. Continued from the previous part using the data from your smoking survey, Let's practice the following exercise.

1. Calculate the selected statistical methods for describing the study sample using your own data and present in the table given below. Then write in your own word describing the table as well as provide appropriate title for the table.

.....
.....
.....
.....
.....

Table 1. …….

| Characteristics | Number | Percent |
|---|---|---|
| **1. Sex** | | |
| Male | | |
| Female | | |
| **Total** | | |
| **2. Age** | | |
| Less than 18 years | | |
| 18 - 25 years | | |
| 26 years or more | | |
| **Total** | | |
| Mean (Standard Deviation) | | |
| Median (Minimum: Maximum) | | |

1.1 The style for presentation shown above is commonly found in literature and highly recommended. For categorical variable such as sex, the analysis is straight forward. For age which is the continuous variable, we categorized it according to some meaningful cut point then calculate percentage. Additionally we provide five summaries - mean, SD, median, minimum, and maximum. Please comment on this style.

.....
.....
.....

1.2 What is the purpose of examining the distribution of the data since even age is highly skewed but we still report all summary measures.

.....

.....

1.3 Should we add to the table some rows for "Number of cigarette smoked per day"? Justify your answer.

.....

.....

2. Pretending the smoking survey aims to compare the prevalence of smoking between the two communities, one is the community under the smoking cessation promotion project and another is the community without the program. Outline the dummy table for descriptive purpose similar to the above table.

# **PART 5:** STATISTICAL INFERENCE FOR ANSWERING THE RESEARCH QUESTION

1. Sampling variation

1.1 Write down the mean number of cigarettes per day you have obtained from your smoking survey. Then take that figure from all of your class mate.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

1.2 If someone conclude their findings based solely on the statistic, what kind of comments you will make. ……
……
……

1.3 Construct a histogram of the mean number of cigarettes per day superimposed by a frequency curve

1.4 Estimate the mean of the sample mean number of cigarettes per day from the graph shown above = .....
Describe how did you guess that amount .....

1.5 The mean number of cigarettes per day obtained from all 115 subjects was ..... (See the Appendix 1.)

1.6 Let's summarize:

1.6.1 From the data of "n" of 10, we get the sample mean. The mean here is called the "statistics" denoted by $\overline{X}$ , pronounced x-bar.

1.6.2 From the data of "N" of 115, we get the population mean. The mean here is called the "parameter" denoted by μ, pronounced mu.

1.6.3 From the repeated sample, we get the mean of the sample mean which is equivalent to the population mean.

1.6.4 What we aim to estimate is the ***parameter***, i.e., the population mean in this example. In reality, we cannot do like #16.2 for several reasons and thus, almost all situations, the parameter is unknown. The repeated sample (in #16.3) told us that there is the way to get the parameter without having been collected the data from all members of the population. But how? No body did like #16.3 too. We did a study only once. What we have at hand now is only the ***statistic***, i.e., the sample mean (in #16.1) which we are ***not sure*** at all whether or not it closes to the parameter. Thus we need to quantify the "***not sure***" so that we can tell about the parameter using the statistic in terms of "***how many percent sure***". The *percentage* and *probability* in this sense are closely related.

2. Try to understand the Central Limit Theorem

2.1 Compare the distribution of the three types of data from the smoking survey.

**Fig. 1**. Distribution of the data from all members of the population (N = 115).

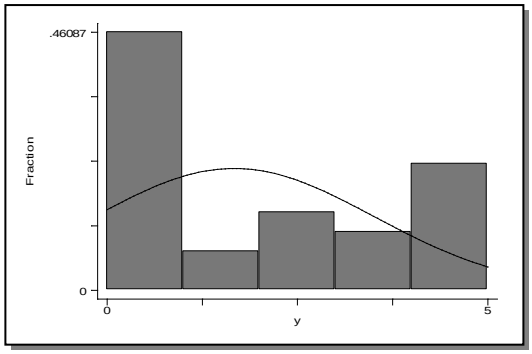
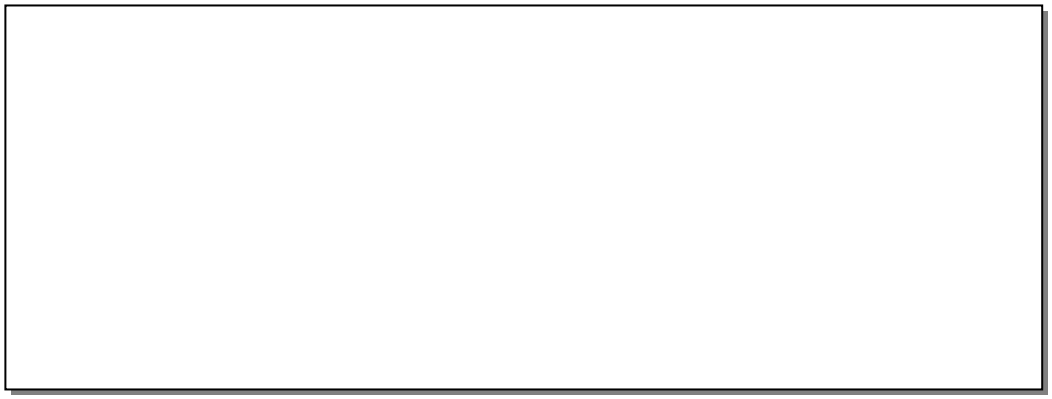


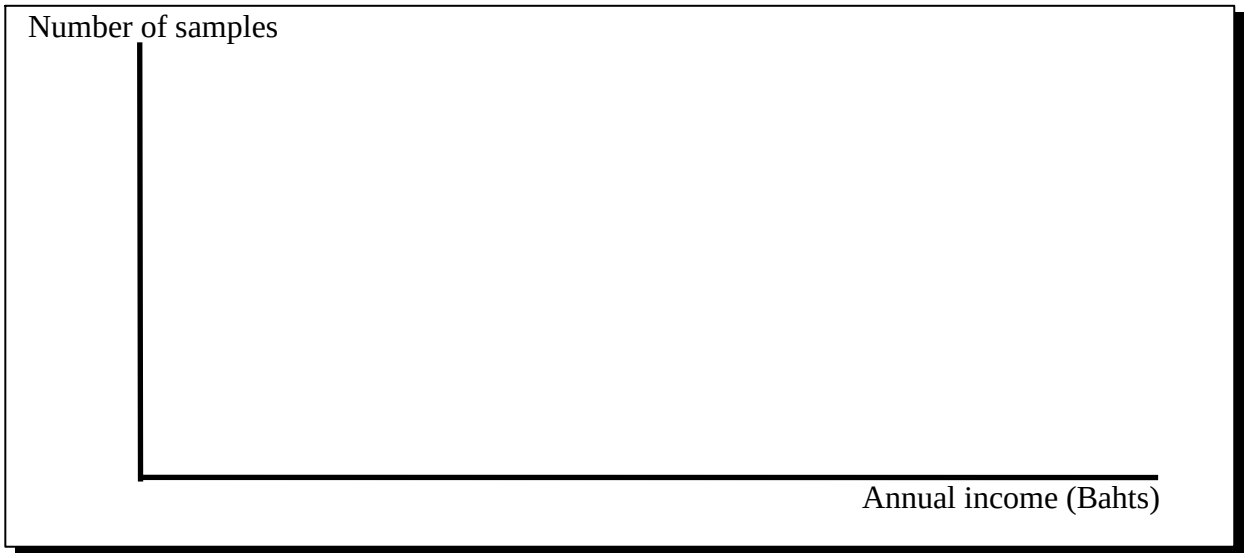

**Fig. 2**. Distribution of the data from a sample of size n = 10 (duplicated from PART 2. Number 2).

**Fig. 3**. Distribution of the data from repeated sample of size n = 10 (duplicated from #1.3 ).



Comments: ……
……
……
……

2.2 From the information in Part 3, Number 4.1, reguarding the hypothetical data that ***"A study of annual income per capita among Thai people found that 80% were poor (poor is classified as those whose income of less than 20000 bahts per year). The range is as small as 500 bahts to several thousand million bahts ".*** If several researchers doing the same study. The mean income they obtained from their study can be used to construct a frequency curve as the bellow figure.




Describe in the probability words about a guy who earn 50,000 Bahts annually:
……
……

2.3 Back to the smoking survey, what shape of the distribution would be if all researchers increase the sample size to 50?

The middle of the curve, i.e. the mean, is …… (the same / smaller / larger).

The base of the curve is …… (the same / narrower / wider).

2.4 We have seen from #2.3 that the width of the base of the curve is affected by sample size. Such width quantify the spread or the deviation of the data. Let's calculate the standard deviation of your smoking survey data. Firstly calculate the mean. Secondly subtract each data from the mean, then square this amount. Thirdly sum all the square of the difference between the data and its mean. Forthly divided the sum by the sample size. This amount is called "variance". Square root of this amount is the standard deviation. We can write in formula as:

$$SD = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{n - 1}}$$

Therefore, your SD is ……
This is the SD of your sample size of 10.

2.5 Do the same for SD of sample mean using the data shown in #1.1.
Therefore, the SD of the sample mean is ……..……
This is called the standard error, denoting SE.

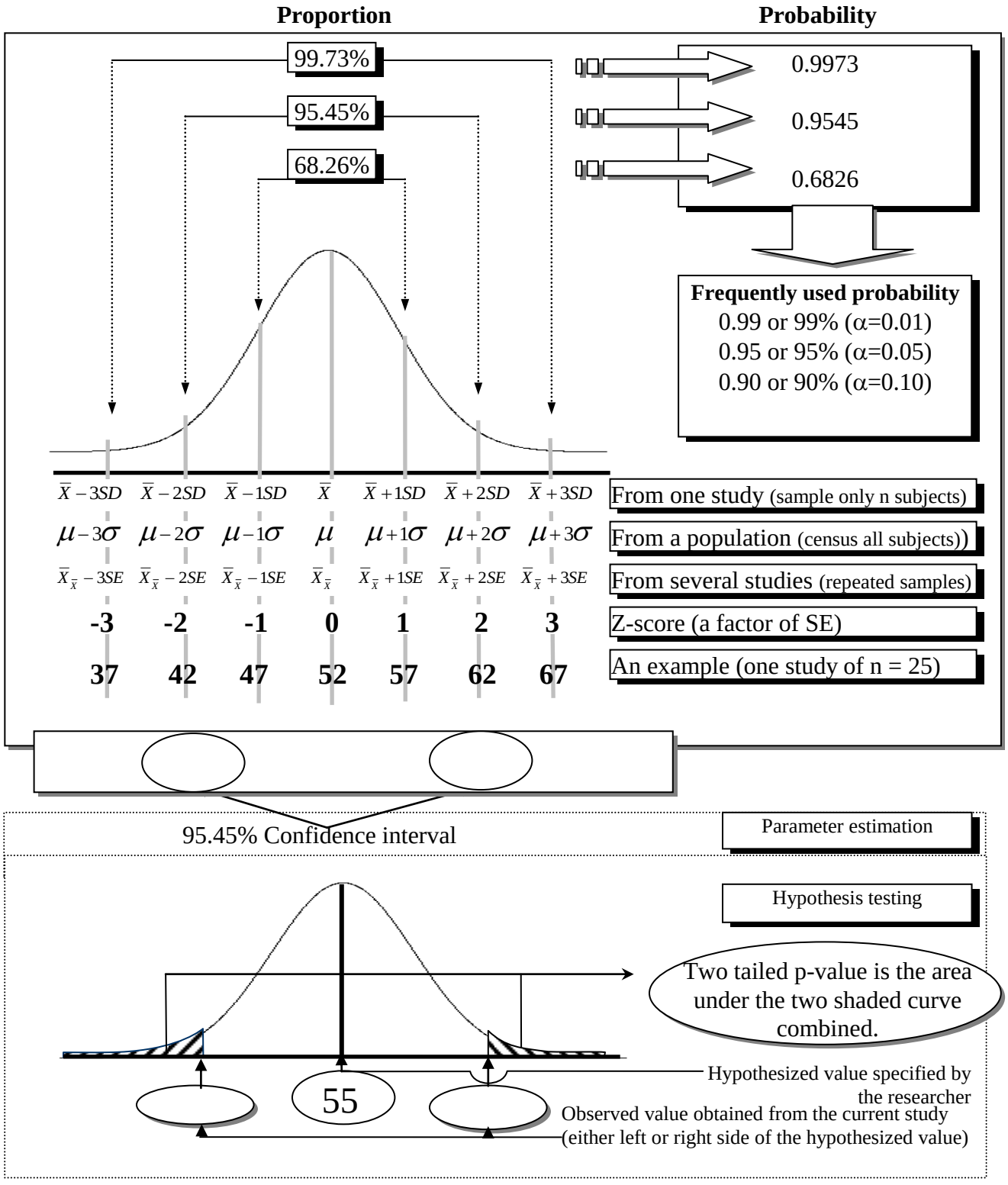
2.6 Let's divided SD obtained in #2.4 by square root of n, you get …….

2.7 Thus SE is approximately equal to SD divided by square root of n.

2.9 Let's summarize:
We have known that we can estimate the population mean without having been collected the data from all members of the population, by estimating it from the sampling distribution obtained from the repeated samples. But we still no hope since nobody does the repeated sample. Then the central limit theorem told us that whatever the distribution of the population data, the sampling distribution always become symmetry. It became the "***normal distribution***" if the sample size is sufficiently large. Knowing the distribution is normal, we can make use of its properties regarding the area under the curve and the score as we have done previously. We have known that the curve can be fully describe if we know its mean and SD. Since the SD of the repeated samples, i.e. the SE can be estimated from the SD that we only have at hand as if we did the repeated sample. Playing around the distribution of the sampling distribution can lead to knowing the population parameter.

3. Make use of the property of the normal distribution.

# Area under the Normal Distribution Curve and other related parameters



Number 3.1 to 3.5 uses the example at the last line of the diagram.

3.1 A study yields a mean score of 52. The SD of this data is ……..

3.2 The SE for this study is ……

3.3 The scores that most likely to include 95.45% of the data in this sample is between …… and ……

*Note that this has nothing to do with the statistical inference, just illustrate its use.*

3.4 The scores that most likely to include 95.45% of the data in the population from which the sample of this study were drawn is between …… and ……

*Note that the above answer is a 95.45% Confidence interval.*

3.5 If the true mean score is 55, what is the probability that a study will yield a mean of 52 or more extreme ……

*Note that the above answer is the p-value.*

3.6 If the sample size is small, the distribution is known not the normal distribution. The base of the curve must be ……. (wider / narrower). It is called the "t - distribution". It will approximate the normal distribution if the sample size is sufficiently large. Thus the shape of the t-distribution depend on the sample size, denoting the degree of freedom (df). For the study that involve one group, the df = n-1. Another popular distribution is the Chi-square distribution. Like the t-distribution, it depend on df and approximate normal distribution for large sample.

4. Summary of general formula for statistical inference

*Confidence interval = Statistic calculated from the study sample ± [(Coefficient × (SE of the statistic]*

$$\textit{Statistical test} = \frac{\textit{Statistic calculated from the study sample - Null value specified under } H_O}{\textit{SE of the statistic}}$$

5. The value of α and the coefficient frequently used for Normal distribution

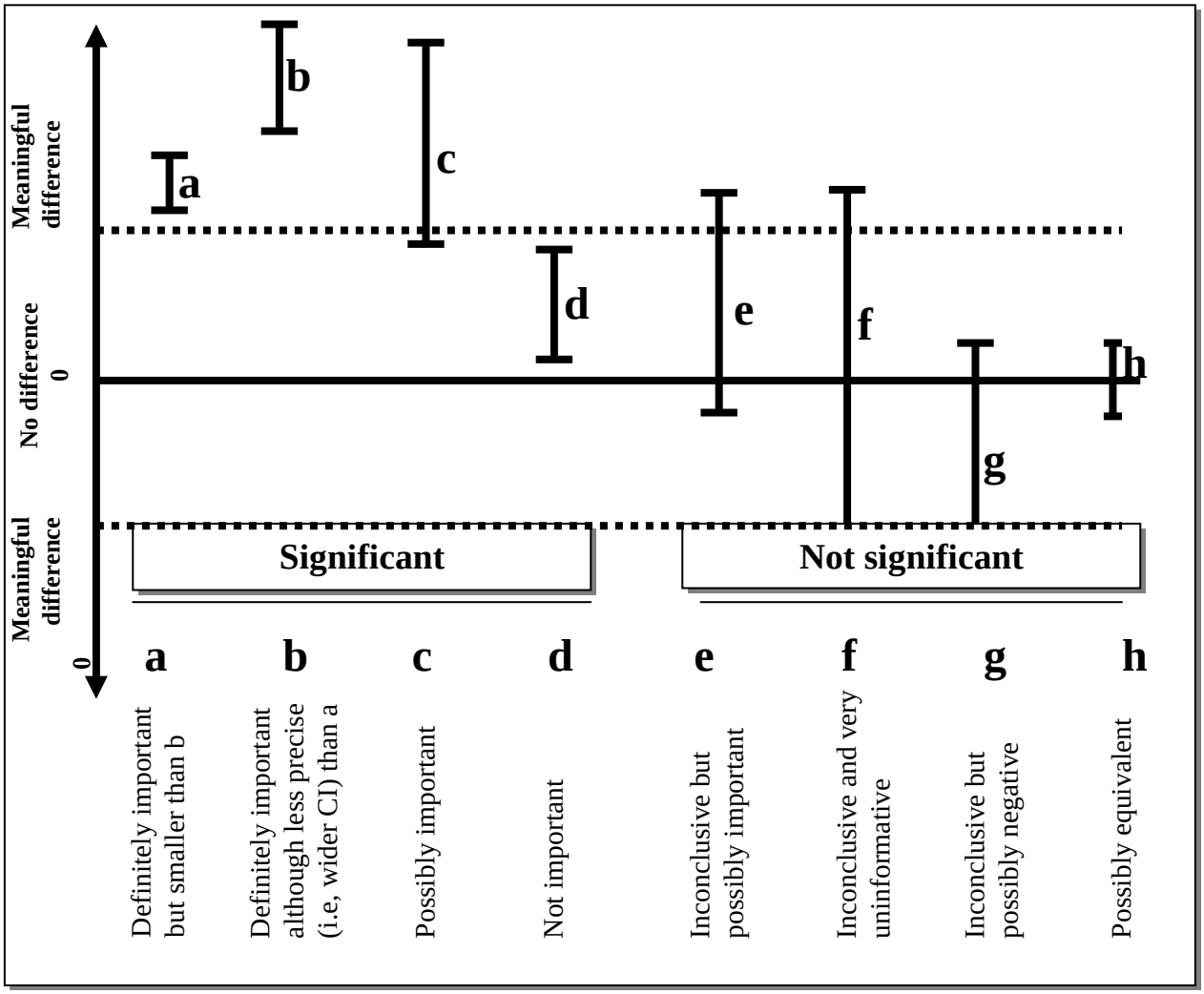
**Values that frequently used**

| **Level of significance** $(\alpha)$ | **Level of confidence** $(1-\alpha)$ | **Coefficient (Z)** |
|---|---|---|
| **0.10** | **0.90** | **1.64** |
| **0.05** | **0.95** | **1.96** |
| **0.01** | **0.99** | **2.58** |

# **PART 6:** CONFIDENCE INTERVAL VS. P-VALUE

Confidence intervals showing eight possible interpretations in terms of statistical significance and practical importance.

*(Adapted from: Armitage, P. and Berry, G. Statistical methods in medical research. 3rd edition. Blackwell Scientific Publications, Oxford. 1994.page 99)*



# **PART 7:** COMPREHENSIVE EXERCISE

1. Complete the following questionnaire.

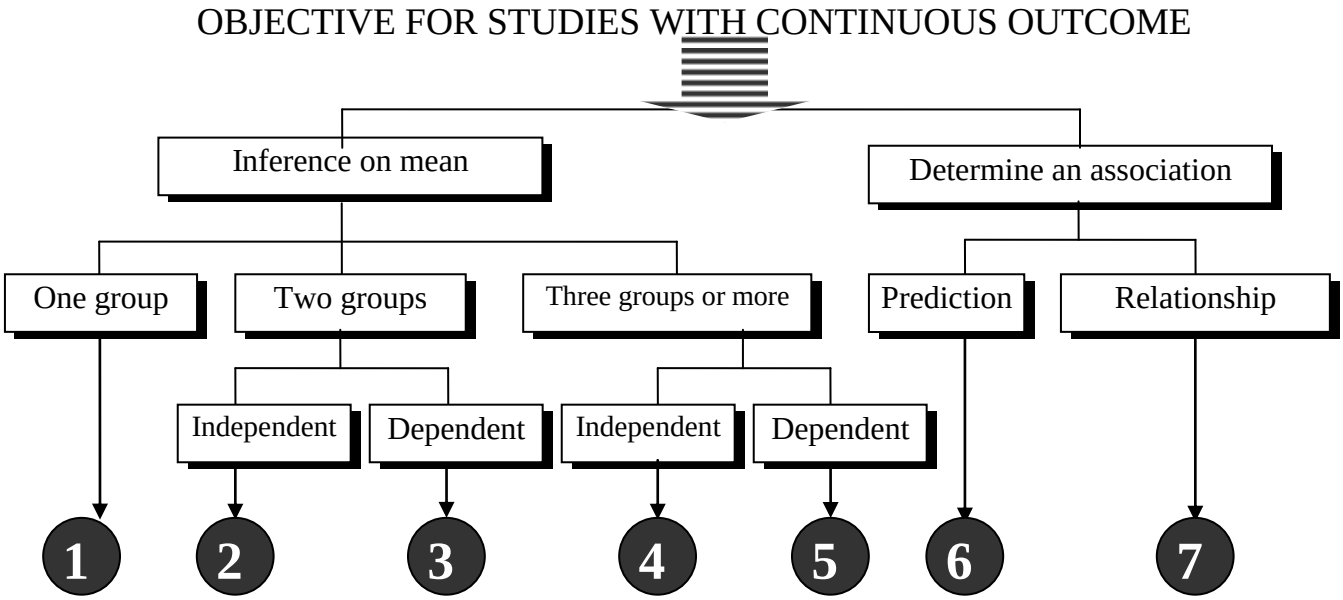
ID.....

## PARTICIPANT HEALTH SURVEY

Please complete the questionnaire by marking 3in [ ] that corresponding to your answer under the first column. For the questions followed by the blank line, please write down your responses.

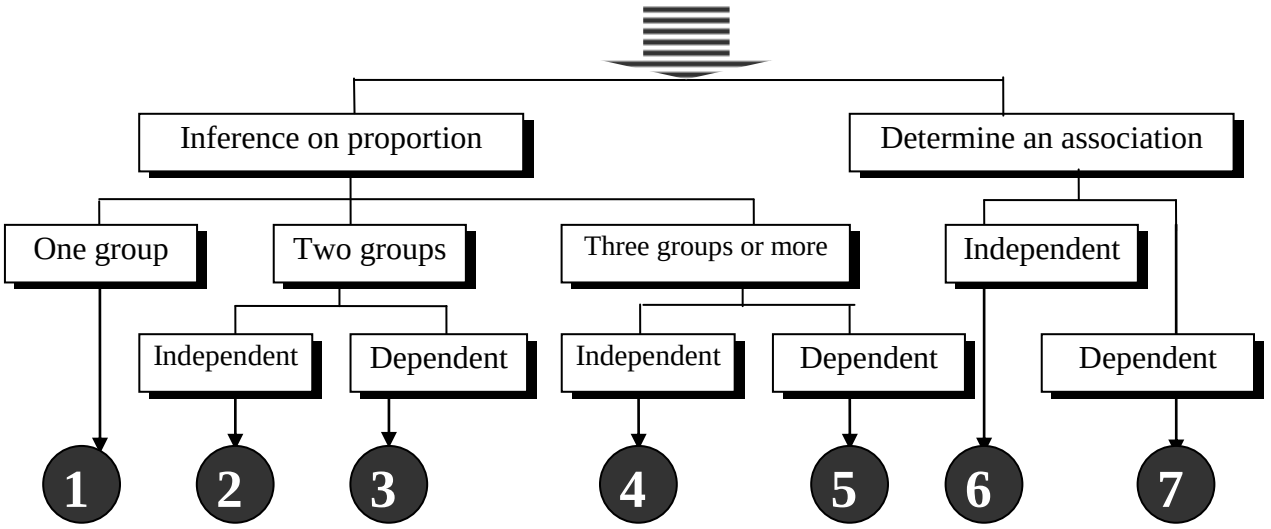
| **Questions** | **For staff only** |
|---|---|
| 1. Name ..... | |
| 2. Sex [ ]1. Male [ ]2. Female | V2[ ] |
| 3. Age ..... years | V3[ ][ ] |
| 4. Height ..... centimeters | V4[ ][ ][ ].[ ] |
| 5. Weight ..... kilograms | V5[ ][ ][ ].[ ] |
| 6. Pulse rate before running in place : ..... | V6[ ][ ][ ] |
| 7. Running in place as fast as possible for 1 minute and then measure your pulse rate: ..... | V7[ ][ ][ ] |
| 8. Rate you mental health status before entering the current course.<br>[ ]1. Stress [ ]2. Not stress | V8[ ] |
| 9. Rate you mental health status at present.<br>[ ]1. Stress [ ]2. Not stress | V9[ ] |
| 10. Rate your overall physical health status<br>[ ]1. Bad [ ]2. Moderate [ ]3. Good [ ]4. Excellent | V10 [ ] |
| 11. What did you do for your last illness<br>[ ]1. Did nothing<br>[ ]2. Self medication<br>[ ]3. Went to a private clinics / hospital<br>[ ]4. Went to public hospital / health center | V11[ ] |

2. Followings are diagrams that are similar to Diagrams 2 and 3 shown earlier in Part 2. For each situation indicated by a number, you are required to state a study objective that could fit to it using the data obtained from the participant health survey.




| |
|---|
| 1. |
| 2. |
| 3. |
| 4. |
| 5. |
| 6. |
| 7. |

OBJECTIVE FOR STUDIES WITH CATEGORICAL OUTCOME




| |
|---|
| 1. |
| 2. |
| 3. |
| 4. |
| 5. |
| 6. |
| 7. |

3. The following examples were adapted from many of published articles. Since these will be used for discussion of appropriateness of statistical methods used, the sources were omitted. Large portion of those were modified for the purpose of training.

   The following examples will be used as the starting point of discussion. The main aim is to Participants are encourage to actively participate in the discussion, raise questions from real practice, and express opinion.

**3.1 Example 1**

In the results section of a research report, there is a sub-section of "Demographic characteristics of the study group" as the following:

***"… Among a total of 517 study subjects, more than three forth were female (76.35%). The mean age is 43 ± 33 years old. Their occupation were farmers, labors, students, government official, and others for 54.01%, 25.03%, 12.16%, 8.00%, and 0.80% respectively (Table 1)."***

**3.2 Example 2**

In the results section of a research report, there is a sub-section of "Comparing baseline characteristics of the study groups" as the following:

***"… Among a total of 211 intervention group and 199 controlled group, most of their characteristics were comparable (non-significant different) except age where there was significantly older in controlled than in intervention groups (p-value < 0.05)."***

**3.3 Example 3**

A conclusion appeared in the abstract of a research is as the following:

***"… To assess the efficacy of ginger in curing acne as compared to usual face washing using ordinary soap, we study 25 school children in each group. The difference between the cure rate were 28% and 36% respectively. This difference was not significant (P > 0.05). We concluded that ginger offer no advantage in such treatment."***

**3.4 Example 4**

In the results section of a research report, there is a sub-section of "Main findings" as the following:

***"… Among a total of 822 exposed group, 5.5% were ill while among a total of 4 who did not expose, 50.00% of them were ill. This finding shows statistically significant (p-value = 0.0033)".***

# **PART 8:** POST-TEST

**A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo (p-value < 0.05).**

*Which of the following statement do you prefer?*

[ ]1. It has been proved that treatment is better than placebo.

[ ]2. If the treatment is not effective, there is less than a 5 percent chance of obtaining such results.

[ ]3. The observed effect of the treatment is so large that there is less than a 5 percent chance that the treatment is no better than placebo.

[ ]4. I do not really know what a p-value is and do not want to guess.

***Source:*** *Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987) What do doctors know about statistics?. Statistics in Medicine, **6**, 3-10.*

Justifications of the answer:

Report the findings of such controlled trial in your own word (*Hints: make up any numbers you need for the report*):

# SUGGESTED READINGS:

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Everitt. B.S. (1994). *Statistical methods for medical investigations*. 3rd edition. London: Edward Arnold.