EXPLORING DATA WITH HISTOGRAM

By

DEMUDU NAGANAIDU

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

August 2020

DEDICATIONS

To all of my love;
Shinu & Puran
Mother & Brothers and Sisters

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

EXPLORING DATA WITH HISTOGRAM

By

DEMUDU NAGANAIDU

August 2020

Chairman : **Associate Professor Dr. Mohd Bakri Bin Adam, Ph.D.**
Faculty : **Institute For Mathematical Research**

Histogram is bla bla bla...

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

EKSPLORASI DATA DENGAN HISTOGRAM

Oleh

DEMUDU NAGANAIDU

Ogos 2020

Pengerusi : **Professor Madya Dr. Mohd Bakri Bin Adam, Ph.D.**
Fakulti : **Institut Penyelidikan Matematik**

Histogram is bla bla bla...

ACKNOWLEDGEMENTS

Histogram is bla bla bla...

THESIS EXAMINATION COMMITTEE

I certify that a Thesis Examination Committee has met on XX August 2020 to conduct the final examination of Demudu Naganaidu on his thesis entitled "Exploring Data with Histogram" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Professor Dr. E, Ph.D.
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairperson)

Professor Dr. F, Ph.D.
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Associate Professor Dr. G, Ph.D.
Professor
School of Mathematical Sciences
Universiti Sains Malaysia
(External Examiner)

Dr. H, Ph.D.
Professor
Laboratoire de Mathématiques Nicolas Oresme
Université de Caen Basse Normandie
France
(External Examiner)

ZULKARNAIN ZAINAL, Ph.D.
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

SUPERVISORY COMMITTEE

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

Associate Professor Dr. Mohd Bakri Bin Adam, Ph.D.
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairperson)

Associate Professor Dr. Jayanthi a/p Arasan, Ph.D.
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Dr. Iskandar Bin Ishak, Ph.D.
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG KIM HUAT, Ph.D.
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

STUDENT DECLARATION

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:_____ Date:_____

Name and Matric No: Demudu Naganaidu, GS49320

SUPERVISORY COMMITTEE DECLARATION

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____
Name of Chairman of Supervisory Committee
Associate Professor Dr. Mohd Bakri Bin Adam, Ph.D.

Signature: _____
Name of Member of Supervisory Committee
Associate Professor Dr. Jayanthi a/p Arasan, Ph.D.

Signature: _____
Name of Member of Supervisory Committee
Dr. Iskandar Bin Ishak, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF ABBREVIATIONS

BFHP	Bivariate Function Hard Problem
CCA	Chosen Ciphertext Attack
CCA1	Non-adaptive CCA
CCA2	Adaptive CCA
CPA	Chosen Plaintext Attack
CRT	Chinese Remainder Theorem
gcd	Greatest Common Divisor
IFP	Integer Factorization Problem
IND	Indistinguishability
IND-CCA2	Indistinguishable against CCA2
LLL	Lenstra-Lenstra-Lovasz
OAEP	Optimal Asymmetric Encryption Padding
ROM	Random Oracle Model
RSA	Rivest-Shamir-Adleman
SET	Secure Electronic Transaction
spm	Single-precision Multiplication

CHAPTER 1

INTRODUCTION

1.1 Background

We are living in a world where technology advancement as continuous phenomena: new applications and devices introduced every now and then such as social media, internet banking, e-commerce, massive open online courses (mooc) , smart-phone, tab, ipad, wearable sensors, self-driving car and etc. These new applications and devices which are important aspect of our daily activities creates massive data (Torrecilla and Romo, 2018). Learning from these data is important for decision making and to improve the process and productivity.

However raw data are often large, making learning from these are almost impossible. Data must be organized, summarized and described in the form that facilitate us to gain information and knowledge. Descriptive statistics which is one of the major two branches of statistic is commonly used to explore the data before any statistical test performed known as inferential statistics.

Exploratory Data Analysis (EDA) is one of the method used in descriptive statistics. EDA was introduced by John W. Tukey in 1977 (Wendy and Angel, 2002). EDA is mostly involves statistical graphs, is an approach of analyzing data visually without a prior assumptions on parametric model of the data, error terms, outliers in data, modality of the data and relationship with other variables (Velleman and Hoaglin, 1981; Wendy and Angel, 2002). Visualising data is vital in exploring data (Scott and Sain, 2005) to help in finding the patterns, structure in data and relationships between variables in data (Hand et al., 2001). EDA helps the researchers to model data based on the what is revealed through exploring with various graphical methods. Generally after the EDA process one would use confirmatory analysis such as hypothesis testing, ANOVA, regression and etc.

Histogram is one of the important tools used in EDA. It is used often in graphical analysis to visualise data (Li et al., 2016). The subject of histogram is so important that it is included in every elementary statistics or data analysis course.

1.2 Histogram

Histogram useful in summarising large amount numerical data graphically (Kirschenmann et al., 2015) and to visualise the distribution of data without prior assumptions on the characteristics of data or the underlying true distribution of the data. Hence histogram is called as a non-parametric density estimator(Keen, 2010). Once a histogram is constructed the general attributes of the data such as symmetry, modality, central location and spread of the data will be revealed.

A typical histogram is a stack of rectangular columns each of whose heights is proportional to the corresponding frequency of data values. Data values grouped into intervals, called as bins (or classes). In constructing histogram number bins or equally the bin width must be decided upfront. Bins can be either same size (i.e equal width) or different size (i.e unequal width) provided it should encompass all the data and non overlap with each other. Number

of bins used to construct histogram will determines the shape of it. Too many bin makes the histogram uneven and unable to find the underlying trend of the data. Too few bin gives little information about the data.

1.3 Histogram Bins

Number of bins in grouping data can be predefined upfront or using some scientific rules. For example data on marks in a exam can be grouped into 5 bins: 0-20, 20 - 40, 40-60, 60-80 and 80-100. Alternatively one can use rules such Sturges(1926) rule to decide number of bins. For sample data of size 40, the Sturges(1926) rule suggest 7 bins. More discussion on rules in Chapter 2 - Literature Review.

1.4 Numerical Data

Numerical data can be classified into discrete and continuous data.

A set of data is said to be discrete if the values belonging to the data set only assumes integer values as 0,1,2,3,4 and etc. There are gaps between two possible values in a discrete dataset. Examples of discrete data are number of children in a family, number of cars on the road, number of students in a school, number of defective items and etc.

Continuous data assumes any values between two specific values. Continuous data are normally obtained by measuring. There are no gaps in between any two possible values in a continuous dataset. Examples of continuous data are height of students in an university, amount of rain in a month, volume of water (in ml) in a bottle, and etc.

1.5 Raw Data and Grouped Data

Dataset can be presented as raw (ungrouped) or grouped data. Raw data is usually unsorted are given as list of data values in the dataset. Grouped data is the raw data that has been sorted into a set classes or intervals or bins. While discrete data can be summarised using a frequency table continuous data must

1.6 Frequency

Frequency is simply counts of a data value appears in a given dataset. Frequency of a dataset displayed using a table format known as frequency table based on the classes or bins. Obtaining frequency of dataset is the first step in summarising dataset before producing graphical charts for visualisation.

Discrete data

1.7 Frequency Table

1.8 Frequency Polygon

1.9 Bar Chart

1.10 Use of Histogram

Histogram can provide a clue to researchers of a possible parametric model suitable for data (Wendy and Angel, 2002) besides detecting unusual observations or behavior in data. Histogram can be used to estimate the 5 number summary of data and variation in data.

Histogram is widely used as one of the Seven (7) Quality Control Tools (7QC) to ensure manufacturing process are within the controlled specification or to maintain the quality level (Magar et al., 2014).

Histogram based image processing that is very useful for medical image analysis, industrial X-ray and etc is used widely because easy to understand and implement besides being effective (Sharma and Kumawat, 2015).

Regression with Histogram Data.

1.11 Problem Statement

Deciding how many bins to be used is still a challenge to most researchers especially when background of the data is unknown. Despite there many suggested methods to determine the number of bins, there is no single method is agreed upon. The five number summary and variation calculated from raw data can used in deciding the number bins.

Detecting the outliers from histogram not very helpful due the method of constructing. The modification of construction histogram can reveal the outliers more effectively.

1.12 Research Aims and Objectives

This study aims to achieve following objectives:

- propose a method using raw data five number summary and variation to decide the number of bins.
- evaluate new method with existing methods
- propose a new method for construction of histogram to detect outliers
- evaluate new method against box-plot method to detect outliers
- propose a fixed frequency histogram for the purpose of segmentation

1.13 Limitation of Study

1.14 Structure of Thesis

CHAPTER 2

LITERATURE REVIEW

2.1 History of Histogram

The term ”histogram” was introduced by Karl Pearson (Ross, 2010; Rufilanchas, 2017) but there was confusion on the date or year it was introduced (Rufilanchas, 2017; Ioannidis, 2003). Several authors stated the year as 1895 (Beniger and Robyn, 1976; Dodge, 2008; Ross, 2010; Li et al., 2016). However (Magnello, 2014; Rufilanchas, 2017) clarified that the histogram was first introduced in 1891 in Gresham Lectures. Karl Pearson used the term while lecturing on ”Maps and Chartograms” (Rufilanchas, 2017).

Beside confusion on year of the term ”histogram” was introduced there is also some confusion on the meaning of ”histogram”. The term is frequently misunderstood that it is used to study history (Rufilanchas, 2017). The association of histogram as historical diagram was due to the fact that Karl Pearson use time histogram in his lecture to describe the ”time-diagram” but he do not meant histogram used to study the subject ”history” (Rufilanchas, 2017).

Histogram has been used long before it received its name i.e from seventeen century. (Scott, 1979; Loquin and Strauss, 2008) believes that John Graunt (as cited in Westergard , 1968, p.22) probably was using histogram in 1662. John Graunt(1662) used a histogram to describe the mortality rate (Copas and Haberman, 1983; Scott and Sain, 2005). However (Chen et al., 2007; Ross, 2010; Rufilanchas, 2017) believes that William Playfair (1759- 1823) who invented bar chart was the person responsible in introducing histogram in 1801.

Andre-Michael Guerry (A.M Guerry) a French mathematician was also given credit in introducing histogram in 1833 when he use William Playfair’s bar chart concept on continuous data by categorizing i.e data on age and time. (Beniger and Robyn, 1976; Friendly, 2007; Ross, 2010). A.M. Guerry produced his first publication in 1829 about crime in France population. In 1833 he return with his second publication by refining and extending data on his first published work. His third and major work that happened in 1864. Third publication used histogram as one of the main tool in presenting the comparison of crime against population between France and England (Friendly, 2007).

According to (Ross, 2010) the systematic development in histogram was due to the work of Adolphe Quetelet in 1846 where he introduce methodological steps in doing research. Quetelet who was a Belgian statistician gave importance to graphical visualization for social science researches. Since then the histogram was further developed such as comparing data from different categories. For example Francis A. Walker[1840-1897] in producing the Atlas of Ninth Census for U.S nationals in 1874 incorporated a bilateral histogram to show graphically the distribution of deaths by sex and month of death and according to race and nationality (Beniger and Robyn, 1976; Friendly, 2008).

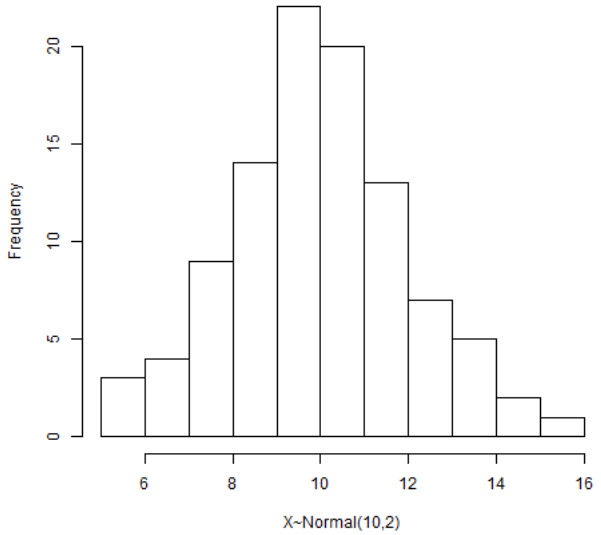
Despite being one of the oldest graphical display tool, histogram is still being relevant for presenting univariate continuous data visually (Scott, 1979; Wand, 1997) and also being used in recent ”Big Data Analytics” (Berger et al., 2016).

CHAPTER 3

METHODOLOGY

The most common histogram found and being used to explore the distribution of data. Each bin in equal width and non overlap.

Figure 3.1: Equal Width Histogram



3.1 Classical Histogram

3.1.1 Flow of Constructing Histogram

1. Selecting Bin Width
2. Selecting Starting Point for First Bin
3. Constructing the frequency Table
4. Constructing the Histogram

1. Selecting Number of Bin
2. Selecting Starting Point for First Bin
3. Constructing the frequency Table
4. Constructing the Histogram

3.1.2 Selection of Bins

3.1.3 Equal Width

3.1.4 Unequal Width

3.1.5 Symmetrical Data

3.1.6 Heavy Tail Data

3.1.7 Big Data and Small Data

3.1.8 Missing Data

Frequency Table

compare the distribution with normal,

3.2 Selecting the Best Histogram Representing Data

3.2.1 Goodness of Fit for Histogram

3.2.2 Statistics from frequency table

3.3 Modification of Constructing Histogram

3.3.1 Modification via Bin Numbers

3.3.2 Modification Via Bin Width

3.3.3 Modification for Heavy Tail Data

3.3.4 Starting point of constructing the histogram

3.4 Handling Big Data with Histogram

3.5 Histogram by Area

3.6 Modification of Histogram

3.6.1 Mode Formula

3.6.2 Mean Formula

3.6.3 Kurtosis

3.7 Histogram by Area Fall Down(Order Frequency Histogram)

3.8 Percentile Histogram

BIBLIOGRAPHY

Beniger, J. R. and Robyn, D. L. (1976). The history and future of graphics in statistics. *Social Statistics Proceedings*, pages 192–197.

Berger, M. A., Mathew, P. A., and Walter, T. (2016). Big data analytics in the building industry. *ASHRAE Journal*, 58(LBNL–1005983).

Chen, C.-h., Härdle, W. K., and Unwin, A. (2007). *Handbook of data visualization*. Springer Science & Business Media.

Copas, J. and Haberman, S. (1983). Non-parametric graduation using kernel methods. *Journal of the Institute of Actuaries (1886-1994)*, 110(1):135–156.

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer Science & Business Media.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

Francis, L. A. (2005). Dancing with dirty data methods for exploring and cleaning data.

Friendly, M. (2007). A.-m. guerry's" moral statistics of france": Challenges for multivariable spatial analysis. *Statistical Science*, pages 368–399.

Friendly, M. (2008). The golden age of statistical graphics. *Statistical Science*, pages 502–535.

Giudici, P. (2005). *Applied data mining: Statistical methods for business and industry*. John Wiley & Sons.

Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning)*. MIT press Cambridge, MA.

Ioannidis, Y. (2003). -the history of histograms (abridged). In *Proceedings 2003 VLDB Conference*, pages 19–30. Elsevier.

John, C., Cleveland, W., Kleiner, B., and Tukey, P. (1983). Graphical methods for data analysis. *Wadsworth, Ohio*, pages 128–129.

Keen, K. J. (2010). *Graphics for statistics and data analysis with R*. CRC Press.

Kirschenmann, T., Damien, P., and Walker, S. (2015). A note on the e–a histogram. *Statistics & Probability Letters*, 103:105–109.

Leek, J. (2015). The elements of data analytic style. *J. Leek.Amazon Digital Services, Inc*.

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. *arXiv preprint arXiv:1612.07216*.

Loquin, K. and Strauss, O. (2008). Histogram density estimators based upon a fuzzy partition. *Statistics & Probability Letters*, 78(13):1863–1868.

Magar, V. M., Shinde, V. B., et al. (2014). Application of 7 quality control (7 qc) tools for continuous improvement of manufacturing processes. *International Journal of Engineering Research and General Science*, 2(4):364–371.

Magnello, E. (2014). *Introducing Statistics: A Graphic Guide*. Icon Books Ltd.

Peng, R. (2012). *Exploratory data analysis with R*. Lulu. com.

Ross, S. M. (2010). Introductory statistics.

Rufilanchas, D. R. (2017). On the origin of karl pearsons term histogram. *Estadística española*, 59(192):29–35.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Scott, D. W. and Sain, S. R. (2005). Multidimensional density estimation. *Handbook of statistics*, 24:229–261.

Sharma, U. K. and Kumawat, K. (2015). Review of histogram based image contrast enhancement techniques. *Int. Journal of Research in Engineering & Technology (IJRET)*, 3(2):65–76.

Torrecilla, J. L. and Romo, J. (2018). Data learning from big data. *Statistics & Probability Letters*.

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Duxbury Press.

Wand, M. (1997). Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–64.

Wendy, L. M. and Angel, R. M. (2002). Computational statistics handbook with matlab. *Chapman& Hall/Crc, NewYork*.

Learning from data is increasingly important in 21st century. We encounter data in our lives on daily basis mostly quantitative data such as prices of goods, weather, exam marks, sales, share prices , account balances, blood pressure reading, height, weight , body mass index and etc. Qualitative data can easily converted to quantitative data by coding.(Alternate Introduction)

Fortunately new fields of study such as Data Analyst, Data Science, Big Data and Data Analytic are also emerged for us to learn from these massive data.

According to Chambers et al., " There is no single statistical tool that is as powerful as a well chosen graph". Often graphical summaries of data are very revealing and helpful in detecting outliers. One of the most commonly used and understood graphical summaries of values of numeric variables is the histogram. Francis (2005).

A histogram is useful to look at when we want to see more detail on the full distribution of the data. The boxplot is quick and handy, but fundamentally only gives us a bit of information. Peng (2012)

Why Graphics? There is no single statistical tool that is as powerful as a well-chosen graph. Our eye-brain system is the most sophisticated information processor ever developed, and through graphical displays we can put this system to good use to obtain deep insight into the structure of data. An enormous amount of quantitative information can be conveyed by graphs; our eye-brain system can summarize vast information qUickly and extract salient features, but it is also capable of focusing on detail. Even for small sets of data, there are many patterns and relationships that are considerably easier to discern in graphical displays than by any other data analytic method.John et al. (1983)

A descriptive data analysis seeks to summarize the measurements in a single data set without further interpretation Leek (2015)

An exploratory data analysis builds on a descriptive analysis by searching for discoveries, trends, correlations, or relationships between the measurements of multiple variables to generate ideas or hypotheses. Leek (2015)

In a quality statistical data analysis the initial step has to be exploratory. This is particularly true of applied data mining, which essentially consists of searching for relationships in the data at hand, not known a priori.Giudici (2005)

More than 50 years ago, John Tukey called for a reformation of academic statistics. In 'The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized science, whose subject of interest was learning from data, or 'data analysis'. Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name Data Science" for his envisioned

field. Donoho (2017)

BIODATA OF STUDENT

The student, **Demudu Naganaidu**, was born in May 1985. Obtained his bachelor degree in Mathematics Education from Universiti Teknologi Malaysia in 2008. He then received a Master of Sciences, from the Institute for Mathematical Research, Universiti Putra Malaysia in 2010. He is currently enrolled a doctoral study in the area of Mathematical Cryptography. He also acts as the secretary of the Malaysian Society of Cryptology Research (MSCR). His research interest involves with designing asymmetric cryptosystems, mathematical cryptanalysis and provable security. The student can be reached via email address; ma_asyraf@upm.edu.my.

LIST OF PUBLICATIONS

The following are the list of publications that arise from this study.

Journal articles:

Muhammad Asyraf Asbullah and Muhammad Rezal Kamel Ariffin (2015). Design of Rabin-like Cryptosystem without Decryption Failure, *Malaysian Journal of Mathematical Sciences* (Accepted for Publication).

Muhammad Asyraf Asbullah and Muhammad Rezal Kamel Ariffin (2015). Design of Rabin-like Cryptosystem without Decryption Failure, *Malaysian Journal of Mathematical Sciences* (Accepted for Publication).

Proceedings:

Muhammad Asyraf Asbullah and Muhammad Rezal Kamel Ariffin (2012). A Proposed CCA-Secure Encryption on an ElGamal Variant. *In the Proceeding of the* 7^{th} *International Conference on Computing and Convergence Technology 2012*, 3 - 5 December 2012, Seoul, pp. 499-503.

UNIVERSITI PUTRA MALAYSIA
STATUS CONFIRMATION FOR THESIS/PROJECT REPORT AND COPYRIGHT
ACADEMIC SESSION: 2016/2017

TITLE OF THE THESIS/PROJECT REPORT:
EXPLORING DATA WITH HISTOGRAM

NAME OF STUDENT: Demudu Naganaidu

I acknowledge that the copyright and other intellectual property in the thesis/project report belonged to Universiti Putra Malaysia and I agree to allow this thesis/project report to be placed at the library under the following terms:

1. This thesis/project report is the property of Universiti Putra Malaysia.
2. The library of Universiti Putra Malaysia has the right to make copies for educational purposes only.
3. The library of Universiti Putra Malaysia is allowed to make copies of this thesis for academic exchange.

I declare that this thesis is classified as:

*Please tick(✓)

☐	CONFIDENTIAL	(contain confidential information under Official Secret Act 1972).
☐	RESTRICTED	(Contains restricted information as specified by the organization/institution where research was done).
☐	OPEN ACCESS	I agree that my thesis/project report to be published as hard copy or online open acces.

This thesis is submitted for:

☐	PATENT	Embargo from _____until _____. (date) (date)

Approved by:

_____
(Signature of Student)
New IC No/Passport No.:850523-10-5567

Date:

_____
(Signature of Chairman of Supervisory Committee)
Name: **Associate Professor Dr. Mohd Bakri Bin Adam, Ph.D.**

Date:

[Note: If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization/institution with period and reasons for confidentially or restricted.]