

สถิติพื้นฐานสำหรับ Machine Learning

Eakasit Pacharawongsakda, Ph.D.

eakasit@datacube.th.ai

Co-Founder and Data Science Team Lead

Cube Analytics Consulting Co., Ltd.

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ชื่อ: เอกสิทธิ์ พชรวงศ์ศักดิ์

- การศึกษา:

- ปริญญาเอก วิทยาการคอมพิวเตอร์
สถาบันเทคโนโลยีนานาชาติสิรินธร (SIIT) มหาวิทยาลัยธรรมศาสตร์
- ปริญญาโท วิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์
- ปริญญาตรี วิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์
(เกียรตินิยมอันดับ 2)

- ประสบการณ์

- Certified RapidMiner Expert and Ambassador
- Research Collaboration with **Western Digital (Thailand)** เฟสที่ 1 ระยะเวลา 6 เดือน
- ร่วมวิจัย โครงการสำรวจข้อมูลเพื่อการวิเคราะห์พฤติกรรมของนักท่องเที่ยวเชิงลึก ด้วยวิธีการทำเหมืองข้อมูล
การท่องเที่ยวแห่งประเทศไทย (ททท)
- วิทยากรอบรมการใช้งานซอฟต์แวร์ open source ทางด้าน data mining



About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



RapidMiner Analyst Certification

This is to Certify that **Eakasit Pacharawongsakda**

Successfully passed the examination for the Certified RapidMiner Analyst.

The RapidMiner Analyst certification level is designed for individuals who wish to demonstrate a fundamental understanding of how RapidMiner software works and is used.

Certified Analyst professionals will be able to prepare data and create predictive models in standard data environments typically found within most analyst positions.

The candidate has proven the ability to:

- Prepare data
- Build predictive models
- Evaluate the model's quality
- Score new data sets
- Deploy data mining models

With:

- RapidMiner Studio
- RapidMiner Server

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



Certification

This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for the

Certified RapidMiner Expert.

The RapidMiner Expert certification level is designed to demonstrate a complete understanding of how RapidMiner Studio and its ecosystem works and is used. Certified professionals will be able to use RapidMiner Studio to analyze Text and Web Information, deploy solutions to the RapidMiner Server, create reports and integrate with Hadoop platforms via RapidMiner Radoop. The candidate has proven the ability to,

- Perform Text and Web Mining
- Deploy and manage a RapidMiner Server
- Create and deploy RapidMiner WebApps
- Develop and integrate RapidMiner dashboard elements
- Connect RapidMiner Studio and Server to a Hadoop platform
- Perform machine learning tasks using the Radoop operators

with

- RapidMiner 8

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Applications & Use Cases Professional Certification

The Professional level is the first of two certification levels and is rewarded for solid experience in the topic. Some of the topics tested at this level are:

Introduction to Machine Learning and Data Science, CRISP-DM and Variants, Use Cases for Machine Learning, Visualization and What to do with Models.

This certificate can be verified using the open badge which was issued to the certificate holder.

The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck




Knut Makowski
(Head of Education)


Peter Lee
(CEO RapidMiner)

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Applications & Use Cases Master Certification

The Master level is the second level after Professional and stands for high proficiency in the topic. Some of the topics tested at this level are:

Running Processes, Deploying Models, Model Management and Web Apps.

This certificate can be verified using the open badge which was issued to the certificate holder.

The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck




Knut Makowski
(Head of Education)


Peter Lee
(CEO RapidMiner)

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Data Engineering Professional Certification

The Professional level is the first of two certification levels and is rewarded for solid experience in the topic. Some of the topics tested at this level are:

Data Access, Basic Transformations, Working with multiple Data Sets, Pivot Tables, Routines and Simple Text Processing.

This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck



Knut Makowski
Knut Makowski
(Head of Education)

Peter Lee
Peter Lee
(CEO RapidMiner)

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Data Engineering Master Certification

The Master level is the second level after Professional and stands for high proficiency in the topic. Some of the topics tested at this level are:

Loops and Branches, Advanced Text Processing, Exception Handling, Logging, Data Cleansing, Regular Expressions, Macros, Web APIs, and Scripting.

This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck



Knut Makowski
Knut Makowski
(Head of Education)

Peter Lee
Peter Lee
(CEO RapidMiner)

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Machine Learning Professional Certification

The Professional level is the first of two certification levels and is rewarded for solid experience in the topic. Some of the topics tested at this level are:

Classification, Regression, Scoring, Hold-out Validation, Correlations, Feature Importance, Clustering and Association Rules.

This certificate can be verified using the open badge which was issued to the certificate holder.

The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck



Knut Makowski
Knut Makowski
(Head of Education)

Peter Lee
Peter Lee
(CEO RapidMiner)

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Machine Learning Master Certification

The Master level is the second level after Professional and stands for high proficiency in the topic. Some of the topics tested at this level are:

Complex Predictive Models, Cross Validation and Correct Validation, Parameter Optimization, Model Selection, Feature Engineering, Time Series and Forecasting and Integrating R and Python models.

This certificate can be verified using the open badge which was issued to the certificate holder.

The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck



Knut Makowski
Knut Makowski
(Head of Education)

Peter Lee
Peter Lee
(CEO RapidMiner)

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning



This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Platform Administration Master Certification

The Master level stands for high proficiency in the topic. Some of the topics tested at this level are:

Platform Overview, RapidMiner Studio Installation, RapidMiner Server Administration, RapidMiner Radoop, RapidMiner Real-Time Scoring, and RapidMiner Marketplace.



This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck


Knut Makowski
(Head of Education)


Peter Lee
(CEO RapidMiner)

About Data Cube

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

Founded on

2013

Focused on training/implementation related to

**Data Science / AI
Machine Learning /
Big Data**

100+

public/in-house
training classes

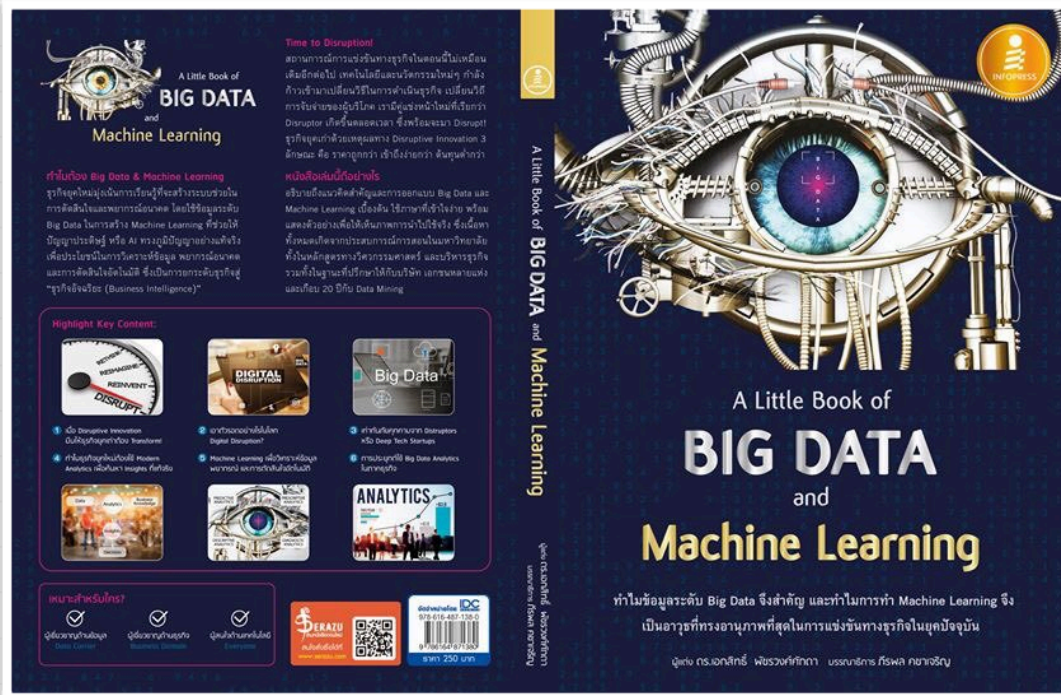
47,000+

people like Facebook page

About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- หนังสือ Big Data และ Machine Learning ฉบับภาษาไทย



About us

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- หนังสือ Data Mining ฉบับภาษาไทย



Basic Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- สถิติ (Statistics) หมายถึง **ตัวเลขที่แสดงข้อเท็จจริงเกี่ยวกับเรื่องใดเรื่องหนึ่ง**
 - สถิติที่แสดงปริมาณน้ำฝน
 - สถิติอุบัติเหตุ
- สถิติ (Statistics) หมายถึง **ศาสตร์ หรือหลักการและระเบียบวิธีทางสถิติที่ว่า**
ด้วย
 - การเก็บรวบรวมข้อมูล
 - การนำเสนอข้อมูล
 - การวิเคราะห์ข้อมูล
 - การตีความหมายข้อมูล

Basic Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- สถิติจะ**แบ่งได้เป็น 2 ประเภทใหญ่ๆ** คือ
 - สถิติพรรณนา (Descriptive Statistics)
 - สถิติเชิงอนุมาน (Inferential Statistics)

Basic Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- **สถิติพรรณนา (Descriptive Statistics)**

- การ**บรรยายลักษณะของข้อมูล** (Data) ที่ถูกเก็บรวบรวมจากประชากรหรือกลุ่มตัวอย่างที่น่าสนใจ ซึ่งอาจจะแสดงในรูป
 - ค่าเฉลี่ย (Mean)
 - ค่ามัธยฐาน (Median)
 - ค่าฐานนิยม (Mode)
 - ส่วนเบี่ยงเบนมาตรฐาน (Deviation)
 - ความแปรปรวน (Variance)

Basic Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- **สถิติเชิงอนุมาน (Inferential Statistics)**

- สถิติที่ว่าด้วย**การวิเคราะห์ข้อมูลที่รวบรวมมา**จากกลุ่มตัวอย่าง เพื่อ**อธิบายสรุปลักษณะบางประการ**ของประชากร โดยมีการนำทฤษฎีความน่าจะเป็นมาประยุกต์ใช้ สถิติสาขานี้ ได้แก่
 - การประมาณค่าทางสถิติ
 - การทดสอบสมมุติฐานทางสถิติ
 - การวิเคราะห์การถดถอยและสหสัมพันธ์

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- การสรุปค่าสำหรับ 1 ตัวแปร มักจะใช้วิธีการวัด การหาค่าเข้าสู่ศูนย์กลาง (measure of central tendency)

ค่าเฉลี่ย (Mean)

ผลรวมของค่าข้อมูล
ทั้งหมดหารด้วยจำนวน
ข้อมูลทั้งหมด

$$\frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$$

ค่ากลางหรือมัธยฐาน (Median)

ค่าที่อยู่ตรงกลางหลัง
จากการเรียงลำดับ
ข้อมูลจากน้อยไปมาก

ค่าฐานนิยม (Mode)

ค่าที่เกิดขึ้นบ่อยที่สุดใน
ชุดข้อมูล

ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

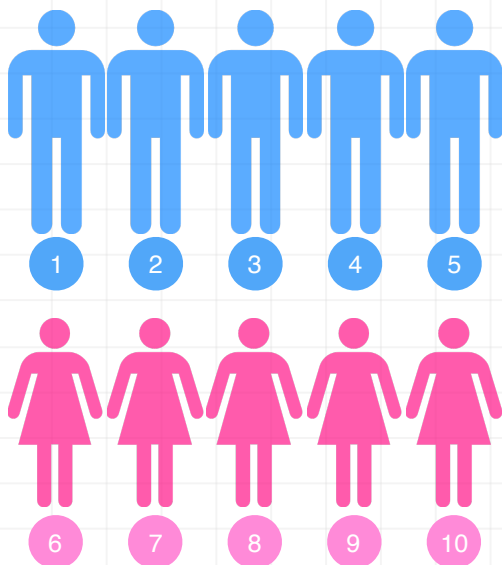
ค่าที่ต่างจากค่าเฉลี่ยเป็น
จำนวนเท่าไร

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots}{N}$$

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ข้อมูลตัวอย่าง



ID	ส่วนสูง	น้ำหนัก
1	170	70
2	170	75
3	180	75
4	185	80
5	170	80
6	150	45
7	155	50
8	150	50
9	160	55
10	165	50

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- การสรุปค่าสำหรับ 1 ตัวแปร มักจะใช้วิธีการวัด การหาค่าเข้าสู่ศูนย์กลาง (measure of central tendency)

ค่าเฉลี่ย (Mean)

ผลรวมของค่าข้อมูล
ทั้งหมดหารด้วยจำนวน
ข้อมูลทั้งหมด

$$\frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$$

ค่ากลางหรือมัธยฐาน (Median)

ค่าที่อยู่ตรงกลางหลัง
จากการเรียงลำดับ
ข้อมูลจากน้อยไปมาก

ค่าฐานนิยม (Mode)

ค่าที่เกิดขึ้นบ่อยที่สุดใน
ชุดข้อมูล

ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

ค่าที่ต่างจากค่าเฉลี่ยเป็น
จำนวนเท่าไร

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots}{N}$$

Descriptive Statistics: Mean

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ID	ส่วนสูง	น้ำหนัก
1	170	70
2	170	75
3	180	75
4	185	80
5	170	80
6	150	45
7	155	50
8	150	50
9	160	55
10	165	50

- ค่าเฉลี่ย (Mean) ของส่วนสูง
 - $(170+170+180+ \dots + 165)/10 = 165.5$
- ค่าเฉลี่ย (Mean) ของน้ำหนัก
 - $(70+75+75+ \dots + 50)/10 = 63$

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- การสรุปค่าสำหรับ 1 ตัวแปร มักจะใช้วิธีการวัด การหาค่าเข้าสู่ศูนย์กลาง (measure of central tendency)

ค่าเฉลี่ย (Mean)

ผลรวมของค่าข้อมูล
ทั้งหมดหารด้วยจำนวน
ข้อมูลทั้งหมด

$$\frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$$

ค่ากลางหรือมัธยฐาน (Median)

ค่าที่อยู่ตรงกลางหลัง
จากการเรียงลำดับ
ข้อมูลจากน้อยไปมาก

ค่าฐานนิยม (Mode)

ค่าที่เกิดขึ้นบ่อยที่สุดใน
ชุดข้อมูล

ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

ค่าที่ต่างจากค่าเฉลี่ยเป็น
จำนวนเท่าไร

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots}{N}$$

Descriptive Statistics: Median

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ลำดับ	ID	ส่วนสูง
1	6	150
2	8	150
3	7	155
4	9	160
5	10	165
6	1	170
7	2	170
8	5	170
9	3	180
10	4	185

ข้อมูลเรียงลำดับตามส่วนสูง
จากน้อยไปมาก

ลำดับ	ID	น้ำหนัก
1	6	45
2	7	50
3	8	50
4	10	50
5	9	55
6	1	70
7	2	75
8	3	75
9	4	80
10	5	80

ข้อมูลเรียงลำดับตามน้ำหนัก
จากน้อยไปมาก

ค่ากลาง (Median)

- เริ่มจากเรียงลำดับข้อมูลจากน้อยสุดไปหามากสุด
- ค่ากลาง คือ ค่าที่อยู่ตรงกลางหลังจากเรียงลำดับได้แล้ว
 - ในกรณีข้อมูลมีจำนวนเป็นเลขคี่ เช่น 3, 5, 7 ค่ากลางคือค่าที่อยู่ตรงกลาง เช่น ข้อมูลลำดับที่ 2, 3 และ 4 ตามลำดับ
 - ในกรณีที่เป็นเลขคู่ เช่น 8 หรือ 10 ต้องหาค่าเฉลี่ยของลำดับข้อมูลที่อยู่ตรงกลาง ได้แก่ ข้อมูลลำดับที่ 4 และ 5 และข้อมูลลำดับที่ 5 และ 6 ตามลำดับ

Descriptive Statistics: Median

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ลำดับ	ID	ส่วนสูง
1	6	150
2	8	150
3	7	155
4	9	160
5	10	165
6	1	170
7	2	170
8	5	170
9	3	180
10	4	185

ข้อมูลเรียงลำดับตามส่วนสูง
จากน้อยไปมาก

ลำดับ	ID	น้ำหนัก
1	6	45
2	7	50
3	8	50
4	10	50
5	9	55
6	1	70
7	2	75
8	3	75
9	4	80
10	5	80

ข้อมูลเรียงลำดับตามน้ำหนัก
จากน้อยไปมาก

• ค่ากลาง (Median) ของส่วนสูง

- ค่าเฉลี่ยของข้อมูลลำดับที่ 5 และ 6
- $(165+170)/2 = 167.5$

• ค่ากลาง (Median) ของน้ำหนัก

- ค่าเฉลี่ยของข้อมูลลำดับที่ 5 และ 6
- $(55+70)/2 = 62.5$

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ID	รายได้ (บาท)
1	15,000
8	18,500
2	22,000
4	25,000
9	25,000
3	30,000
7	30,000
6	35,000
10	35,000
5	45,000
11	10,000,000

• ค่ากลางสำหรับข้อมูลที่เป็นตัวเลขมี 2 ค่า คือ

- ค่าเฉลี่ย (Mean) และ ค่ากลาง หรือ มัธยฐาน (Median)
- ข้อเสียของ **ค่าเฉลี่ย (Mean)** จะอ่อนไหวกับ **ค่าที่เป็น outlier** หรือค่าที่สูงหรือน้อยกว่าปกติ เช่น ในตารางจะได้ ค่าเฉลี่ย (mean) ของรายได้เท่ากับ 934,591 บาท
- แต่ค่า**มัธยฐาน (Median)** จะไม่อ่อนไหวกับ **ค่า Outlier** ดังนั้น ค่ามัธยฐานของรายได้จะเท่ากับ 30,000 บาท ซึ่งเป็นตัวแทนของรายได้คนส่วนใหญ่มากกว่า

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- การสรุปค่าสำหรับ 1 ตัวแปร มักจะใช้วิธีการวัด การหาค่าเข้าสู่ศูนย์กลาง (measure of central tendency)

ค่าเฉลี่ย
(Mean)

ผลรวมของค่าข้อมูล
ทั้งหมดหารด้วยจำนวน
ข้อมูลทั้งหมด

$$\frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$$

ค่ากลางหรือมัธยฐาน
(Median)

ค่าที่อยู่ตรงกลางหลัง
จากการเรียงลำดับ
ข้อมูลจากน้อยไปมาก

ค่าฐานนิยม
(Mode)

ค่าที่เกิดขึ้นบ่อยที่สุดใน
ชุดข้อมูล

ค่าเบี่ยงเบนมาตรฐาน
(Standard Deviation)

ค่าที่ต่างจากค่าเฉลี่ยเป็น
จำนวนเท่าไร

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots}{N}$$

Descriptive Statistics: Mode

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ลำดับ	ID	ส่วนสูง
1	6	150
2	8	150
3	7	155
4	9	160
5	10	165
6	1	170
7	2	170
8	5	170
9	3	180
10	4	185



ส่วนสูง	ความถี่
150	2
155	1
160	1
165	1
170	3
180	1
185	1

- Mode ของส่วนสูง
- ค่าที่เกิดขึ้นบ่อยสุดในส่วนสูง
- = 170

ข้อมูลเรียงลำดับตามส่วนสูง
จากน้อยไปมาก

Descriptive Statistics: Mode

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ลำดับ	ID	น้ำหนัก
1	6	45
2	7	50
3	8	50
4	10	50
5	9	55
6	1	70
7	2	75
8	3	75
9	4	80
10	5	80



น้ำหนัก	ความถี่
45	1
50	3
55	1
70	1
75	2
80	2

- Mode ของส่วนสูง
 - ค่าที่เกิดขึ้นบ่อยสุดในส่วนสูง = 170
- Mode ของส่วนน้ำหนัก
 - ค่าที่เกิดขึ้นบ่อยสุดในส่วนสูง = 50

ข้อมูลเรียงลำดับตามน้ำหนัก
จากน้อยไปมาก

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- การสรุปค่าสำหรับ 1 ตัวแปร มักจะใช้วิธีการวัด การหาค่าเข้าสู่ศูนย์กลาง (measure of central tendency)

ค่าเฉลี่ย
(Mean)

ผลรวมของค่าข้อมูล
ทั้งหมดหารด้วยจำนวน
ข้อมูลทั้งหมด

$$\frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$$

ค่ากลางหรือมัธยฐาน
(Median)

ค่าที่อยู่ตรงกลางหลัง
จากการเรียงลำดับ
ข้อมูลจากน้อยไปมาก

ค่าฐานนิยม
(Mode)

ค่าที่เกิดขึ้นบ่อยที่สุดใน
ชุดข้อมูล

ค่าเบี่ยงเบนมาตรฐาน
(Standard Deviation)

ค่าที่ต่างจากค่าเฉลี่ยเป็น
จำนวนเท่าไร

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots}{N}$$

Descriptive Statistics: STDEV

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ลำดับ	ID	ส่วนสูง	ผลต่าง
1	6	150	150 - 165.5 = -15.5
2	8	150	150 - 165.5 = -15.5
3	7	155	155 - 165.5 = -10.5
4	9	160	160 - 165.5 = -5.5
5	10	165	165 - 165.5 = -0.5
6	1	170	170 - 165.5 = 4.5
7	2	170	170 - 165.5 = 4.5
8	5	170	170 - 165.5 = 4.5
9	3	180	180 - 165.5 = 14.5
10	4	185	185 - 165.5 = 19.5

• Standard Deviation ของส่วนสูง

$$\bullet \sqrt{[(-15.5)^2 + (-15.5)^2 + \dots +]/10}$$

$$\bullet = 1.28$$

ข้อมูลเรียงลำดับตามส่วนสูงจากน้อยไปมาก

Descriptive Statistics: STDEV

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ลำดับ	ID	น้ำหนัก	ผลต่าง
1	6	45	45 - 63 = -18
2	7	50	50 - 63 = -13
3	8	50	50 - 63 = -13
4	10	50	50 - 63 = -13
5	9	55	55 - 63 = -8
6	1	70	70 - 63 = 7
7	2	75	75 - 63 = 12
8	3	75	75 - 63 = 12
9	4	80	80 - 63 = 17
10	5	80	80 - 63 = 17

• Standard Deviation ของส่วนสูง

$$\bullet \sqrt{[(-15.5)^2 + (-15.5)^2 + \dots +]/10}$$

$$\bullet = 1.28$$

• Standard Deviation ของน้ำหนัก

$$\bullet \sqrt{[(-18)^2 + (-13)^2 + \dots + 17^2]/10}$$

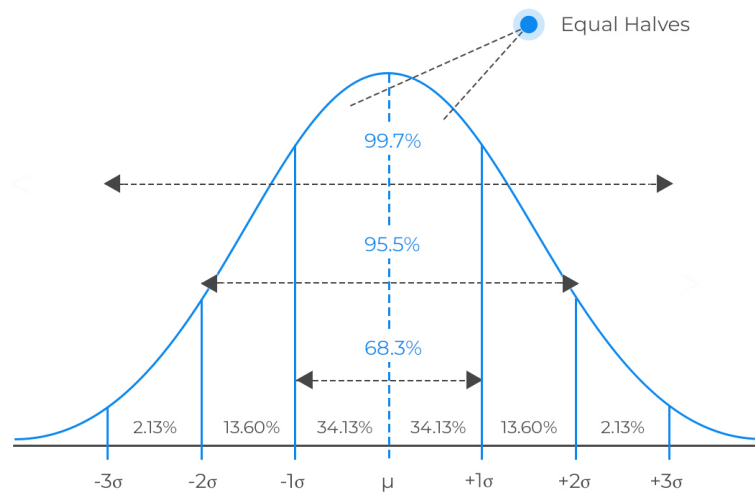
$$\bullet = 3.45$$

ข้อมูลเรียงลำดับตามน้ำหนักจากน้อยไปมาก

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- สามารถ **plot** การกระจายตัวของข้อมูลในรูปแบบของ **curve** ได้โดยทั่วไปเรา
อยากให้อข้อมูลอยู่ใน **รูปแบบของ Normal Distribution** หรือ **รูประฆังคว่ำ**



No. of standard deviations from the mean

image source: <https://analystprep.com/cfa-level-1-exam/quantitative-methods/key-properties-normal-distribution/>

Copyright © 2021 Cube Analytics Consulting Co., Ltd.

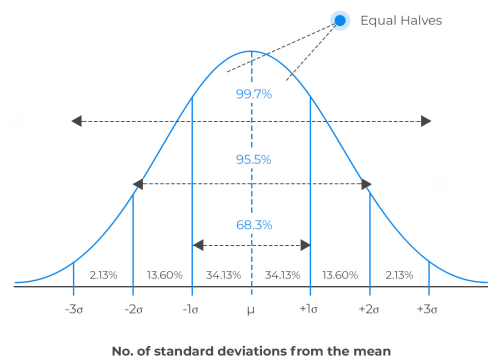
<http://facebook.com/datacube.th>

33

Descriptive Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ถ้าการกระจายตัวของข้อมูลเป็นแบบ Normal Distribution แล้วจะมีคุณสมบัติ
 - มี **การกระจายตัวแบบสมมาตร (symmetric)** โดยมี **ค่าเฉลี่ย (mean)** เป็นตัวแบ่ง
ข้อมูลออกเป็น 2 ส่วน
 - ค่าเฉลี่ย (mean), ค่ามัธยฐาน (median) และค่าฐานนิยม (mode) จะ **อยู่ตำแหน่ง
เดียวกัน**
 - ข้อมูลประมาณ 68% จะอยู่ในช่วงค่าเฉลี่ย
 ± 1 ช่วงเบี่ยงเบนมาตรฐาน
 - ข้อมูลประมาณ 95% จะอยู่ในช่วงค่าเฉลี่ย
 ± 2 ช่วงเบี่ยงเบนมาตรฐาน
 - ข้อมูลประมาณ 99% จะอยู่ในช่วงค่าเฉลี่ย
 ± 3 ช่วงเบี่ยงเบนมาตรฐาน



No. of standard deviations from the mean

image source: <https://analystprep.com/cfa-level-1-exam/quantitative-methods/key-properties-normal-distribution/>

Copyright © 2021 Cube Analytics Consulting Co., Ltd.

<http://facebook.com/datacube.th>

34

Rank-Ordered Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- แสดงการกระจายตัวของข้อมูลโดยเรียงข้อมูลตามลำดับจากน้อยไปมาก
- Percentile หรือ Quartile จะเป็นการบอกว่ามีค่าที่น้อยกว่าค่า ณ ตำแหน่งนี้คิดเป็นกี่ % หรือกี่ส่วน

Percentile	Quartile	Metric
0th	0th	ค่าน้อยสุด (minimum)
25th	1st	จำนวนที่น้อยกว่า 25% จะเรียกว่า 1 st quartile
50th	2nd	จำนวนที่น้อยกว่า 50% จะเรียกว่า 2 nd quartile ซึ่งก็คือค่า median
75th	3rd	จำนวนที่น้อยกว่า 75% จะเรียกว่า 3 rd quartile
100th	4th	ค่ามากที่สุด (maximum)

Rank-Ordered Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ตัวอย่างการหาค่า Quartile

ลำดับ	ID	ส่วนสูง
1	6	150
2	8	150
3	7	155
4	9	160
5	10	165
6	1	170
7	2	170
8	5	170
9	3	180
10	4	185

Percentile	Quartile	Value
0th	0th	150
25th	1st	156.25
50th	2nd	167.5
75th	3rd	170
100th	4th	185

Rank-Ordered Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ตัวอย่างการหาค่า Quartile

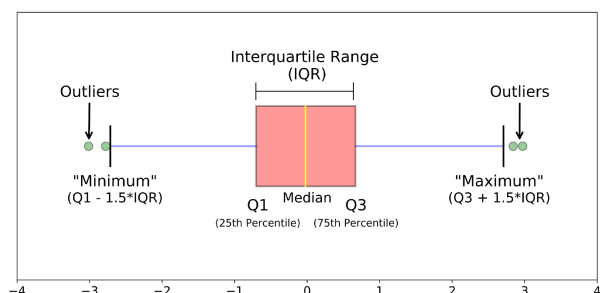
ID	รายได้ (บาท)
1	15,000
8	18,500
2	22,000
4	25,000
9	25,000
3	30,000
7	30,000
6	35,000
10	35,000
5	45,000
11	10,000,000

Percentile	Quartile	Value
0th	0th	15,000
25th	1st	23,500
50th	2nd	30,000
75th	3rd	35,000
100th	4th	10,000,000

Rank-Ordered Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ใช้ในการหา Outlier (ข้อมูลที่ผิดปกติ) ได้
- วิธีการ Inter Quartile Range (IQR)
 - โดยที่ $IQR = Q3 - Q1$
 - ค่าที่เป็น outlier คือ
 - ค่าที่น้อยกว่า $Q1 - 1.5 * IQR$ และ
 - ค่าที่มากกว่า $Q3 + 1.5 * IQR$



Rank-Ordered Statistics

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ID	รายได้ (บาท)
1	15,000
8	18,500
2	22,000
4	25,000
9	25,000
3	30,000
7	30,000
6	35,000
10	35,000
5	45,000
11	10,000,000

- ตัวอย่างการหา Outlier ด้วยวิธีการ IQR

Percentile	Quartile	Value
25th	1 st quartile	23,500
75th	3 rd quartile	35,000

- $IQR = Q3 - Q1 = 35,000 - 23,500 = 11,500$
- ข้อมูลที่เป็น outlier คือจำนวน
 - มีค่าน้อยกว่า $23,500 - 1.5 * 11,500 = 5,750$
 - มีค่ามากกว่า $35,000 + 1.5 * 11,500 = 52,250$
- ดังนั้นข้อมูลหมายเลขที่ 11 ซึ่งมีค่า 10,000,000 ถือว่าเป็น outlier

Correlation

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

“ **Correlation** is a statistic that measures the degree to which two variables move in relation to each other.

If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.”

– Investopedia

(<https://www.investopedia.com/terms/c/correlation.asp>)

“ค่าสหสัมพันธ์เป็นสถิติซึ่งวัดความเกี่ยวข้องกันของตัวแปร 2 ตัวแปรว่าไปในทิศทางเดียวกันหรือไม่ ถ้าไปในทางเดียวกันถือว่ามีความสัมพันธ์เชิงบวก แต่ถ้าไปในทางตรงกันข้ามถือว่ามีความสัมพันธ์เชิงลบ”

Correlation

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- Correlation จะแสดงด้วยค่า Correlation Coefficient หรือค่าสัมประสิทธิ์สหสัมพันธ์
- แสดงได้ดังในสมการ

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

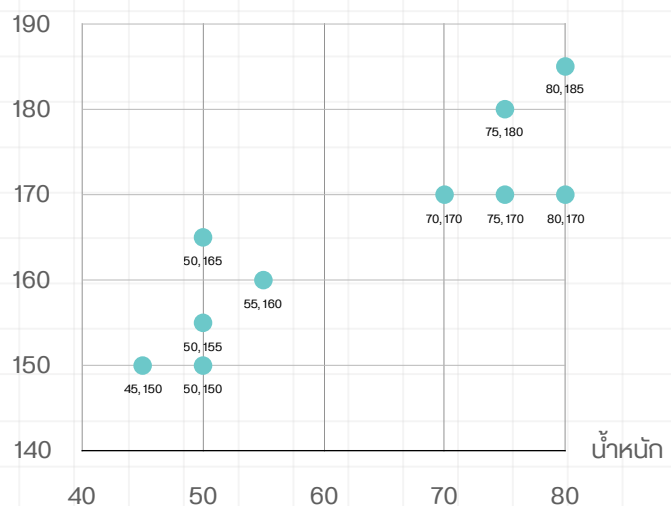
โดยที่

- r_{xy} คือ ค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient)
- x_i และ y_i คือ ข้อมูลของตัวแปร x และตัวแปร y ในลำดับที่ i
- $\bar{x}$ และ $\bar{y}$ คือ ค่าเฉลี่ยของตัวแปร x และตัวแปร y

Correlation

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

ID	ส่วนสูง	น้ำหนัก
1	170	70
2	170	75
3	180	75
4	185	80
5	170	80
6	150	45
7	155	50
8	150	50
9	160	55
10	165	50



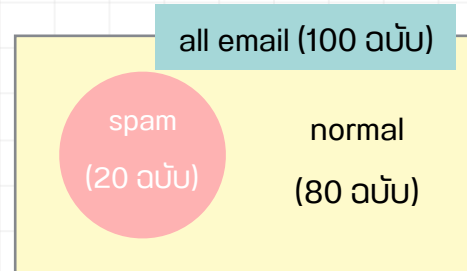
Correlation

0.879

Probability

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- ความน่าจะเป็น (probability)
 - โอกาสที่**เกิดเหตุการณ์จากเหตุการณ์ทั้งหมด** ใช้สัญลักษณ์ $P()$ หรือ $Pr()$
 - โยนเหรียญบาท (มีหัวและก้อย)
 - โอกาสได้หัว มีค่าความน่าจะเป็น $1/2 = 0.5$
 - โอกาสได้ก้อย มีค่าความน่าจะเป็น $1/2 = 0.5$
 - ความน่าจะเป็นของการพบ spam email
 - มี email ทั้งหมด 100 ฉบับ
 - มี spam email ทั้งหมด 20 ฉบับ
 - มี normal email ทั้งหมด 80 ฉบับ
 - โอกาสที่ email จะเป็น spam มีความน่าจะเป็น $20/100 = 0.2$ หรือ $P(\text{spam}) = 0.2$
 - โอกาสที่ email จะเป็น normal มีความน่าจะเป็น $80/100 = 0.8$ หรือ $P(\text{normal}) = 0.8$



Probability

สำหรับการอบรมหลักสูตร Practical Data Science and Machine Learning

- Joint Probability
 - ความน่าจะเป็นที่ 2 เหตุการณ์เกิดร่วมกัน
 - ความน่าจะเป็นที่มีคำว่า Free อยู่ใน spam email
 - สัญลักษณ์ $P(\text{Free}=Y \cap \text{spam})$

