

Data Analytics

ด้วย

สถิติวิเคราะห์เบื้องต้น

พร้อมตัวอย่างการวิเคราะห์ DataInBusiness ด้วยวิธีการทางสถิติ

Data Analytics with Elementary Statistical Analysis

ยวดี เปรมวิชัย ISBN 978-616-582-250-3

279.-

Data Analytics ด้วย สถิติวิเคราะห์เบื้องต้น

โดย ยุวดี เปรมวิชัย

ชื่อเรื่อง Data Analytics ด้วยสถิติวิเคราะห์เบื้องต้น
(Data Analytics by Elementary Statistical Analysis)
ปีที่จัดทำ เมษายน 2564
ชื่อผู้แต่ง ยุวดี เปรมวิชัย e-mail : yuwatisit@gmail.com
ราคา e-book 279.- บาท
ISBN (e-book) 978-616-582-250-3
จำนวนหน้า 168 หน้า
พิสูจน์อักษร/ออกแบบปก/ออกแบบปก/จัดทำโดย ยุวดี เปรมวิชัย

สงวนลิขสิทธิ์ตามพระราชบัญญัติลิขสิทธิ์ 2537

ห้ามผู้ใด คัดลอก ทำซ้ำ ดัดแปลง หรือนำส่วนใดส่วนหนึ่งไปเผยแพร่โดยไม่ได้รับอนุญาตจากเจ้าของผลงาน

คำนำ

หนังสือเล่มนี้เป็นการใช้สถิติวิเคราะห์เบื้องต้น สำหรับ Data Analytics ใน ชั้น Diagnostic Analytics และ Predictive Analytics ประกอบด้วย การทดสอบสมมติฐาน (Hypothesis Testing) และการวิเคราะห์สหสัมพันธ์และการถดถอยเชิงเส้น (Linear Correlation and Regression Analysis) จุดประสงค์เพื่ออธิบายหลักการของสถิติด้วยภาษาง่ายๆ ให้ผู้ปฏิบัติเข้าใจพื้นฐานการวิเคราะห์ทางสถิติ ไม่ลงรายละเอียดของทฤษฎีกล่าวถึงทฤษฎีเพียงเท่าที่จำเป็นสำหรับการปฏิบัติเท่านั้น เน้นการยกตัวอย่างทางการปฏิบัติได้แก่ การจัดการ Data และการใช้สถิติวิเคราะห์เบื้องต้น ทั้งแบบคำนวณเอง และแบบใช้ค่าจากผลลัพธ์ของโปรแกรมสำเร็จรูปทางสถิติ ซึ่งเป็น Tools ที่มีมากมายในปัจจุบัน และถึงอย่างไรพื้นฐานทางสถิติก็ยังคงมีความสำคัญเสมอสำหรับให้เข้าใจผลลัพธ์ของโปรแกรม และมีเหตุผลทางวิชาการในการนำไปใช้ในกิจการจริง สามารถต่อยอดสู่ Machine Learning หรือโมเดลอื่นๆได้

หวังในความเป็นประโยชน์และเพิ่มพูนความรู้ทางสถิติแก่ผู้อ่านทุกท่าน

ยุวดี เปรมวิชัย

กุมภาพันธ์ 2564

สารบัญ

บทที่ 1

หน้า

บทนำ (Introduction)	1
1.1 ความหมายและประเภทของสถิติ	5
1.2 มาตรวัด (Measurement Scale)	8
1.3 คำศัพท์ทางสถิติที่สำคัญ	10

บทที่ 2

หน้า

การทดสอบสมมติฐาน (Hypothesis Testing)	22
2.1 สมมติฐานทางสถิติ (Statistical Hypothesis)	25
2.2 บริเวณวิกฤต (Critical Region)	30
2.3 การเลือกสถิติทดสอบ (Test Statistic)	33
2.4 การสรุปผลการทดสอบ (Conclusion)	43

บทที่ 3

หน้า

การทดสอบสมมติฐานสำหรับ 1 ประชากร (Hypothesis Testing for 1 Population)	48
3.1 การทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ย (Mean) ของ 1 ประชากร	48
3.2 การทดสอบสมมติฐานเกี่ยวกับค่าสัดส่วน (Proportion) ของ 1 ประชากร	58

สารบัญ(ต่อ)

3.3	การทดสอบสมมติฐานเกี่ยวกับค่าความแปรปรวน (Variance)	66
3.4	การทดสอบสมมติฐานโดย Chi-square Test	72

บทที่ 4

หน้า

การทดสอบสมมติฐานสำหรับ 2 ประชากร (Hypothesis Testing for 2 Population)		85
4.1	การทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ยของ 2 ประชากร	86
4.2	การทดสอบสมมติฐานค่าสัดส่วนของ 2 ประชากร	106
4.3	การทดสอบสมมติฐานค่าความแปรปรวนของ 2 ประชากร	112

บทที่ 5

หน้า

การวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA)		120
5.1	การวิเคราะห์ความแปรปรวนทางเดียว (One-way ANOVA)	122
5.2	การทดสอบหลังจากวิเคราะห์ ANOVA (Post Hoc Comparisons)	131

สารบัญ(ต่อ)

บทที่ 6

หน้า

สหสัมพันธ์และการถดถอย (Correlation and Regression)	135
6.1 สหสัมพันธ์ (Correlation)	136
6.2 การถดถอยเชิงเส้น (Linear Regression)	148

บทที่

1

บทนำ
(Introduction)

ปัจจุบันในปี ค.ศ. 2021 นี้ ความรู้เกี่ยวกับ Data เข้ามามีบทบาทในชีวิตประจำวันของมนุษย์อย่างมาก ทำให้พบกับคำว่า Data , Big Data, Data Mining ฯลฯ ซึ่งมีวิธีการเกี่ยวกับการจัดการ Data ด้วยเครื่องมือจัดการ (Data Tools) หลายประเภท หลากหลายรูปแบบ เป็น ตาราง ภาพ กราฟ ฯลฯ จำแนก Data ได้ทุกมิติ เช่น ตารางและกราฟ จำแนกลูกค้าตามเพศ ตามอาชีพ ตามรายได้ เป็นต้น เมื่อจัดการแล้วต้องนำไปใช้ประโยชน์ เช่น บริษัทจำหน่ายสินค้ามี Data ลูกค้าที่ได้จัดการไว้ครบทุกมิติแล้ว ก็ต้องใช้ประโยชน์ เช่น ลูกค้าเพศชายชอบสินค้าแบบใด ลูกค้าหญิงชอบสินค้าแบบใด ซึ่งการใช้ประโยชน์จาก Data ต้องอาศัยการเก็บรวบรวม Data ที่มีมากมาย เอาเฉพาะที่เกี่ยวข้อง นำไปวิเคราะห์เพื่อหาข้อสรุปนำไปสู่การตัดสินใจที่มีหลักวิชาการ วิธีการวิเคราะห์ Data ล้วนเป็นวิธีการทางสถิติ ความรู้ความเข้าใจในสถิติวิเคราะห์ (Statistical Analysis) จึงเป็นพื้นฐานที่สำคัญมากสำหรับวิเคราะห์เพื่อใช้ประโยชน์จาก Data ทำให้เกิดการศึกษาเรียนรู้ที่ปัจจุบันเรียกว่า Data Analytics ขึ้น

Data Analytics หมายถึง วิธีการทางสถิติที่ใช้สำหรับวิเคราะห์ Data เพื่อหาข้อสรุปที่จะนำไปสู่การตัดสินใจประกอบกิจการใดกิจการหนึ่ง โดยเป็นการตัดสินใจที่มีหลักการของวิชาการ ใช้ความรู้ของการวิเคราะห์ทางสถิติ แสดงการตัดสินใจอย่างชัดเจนในกิจการนั้นๆ วิธีการของ Data Analytics จึงแบ่งตามวิธีการวิเคราะห์ ได้ 4 ระดับ ได้แก่

1) Descriptive analytics เป็นการวิเคราะห์ขั้นพื้นฐานเปรียบได้กับการที่พบ Data ครั้งแรก เป็นการวิเคราะห์เชิงบรรยายเพื่ออธิบายข้อมูลนั้น ณ เวลานั้นๆ ไม่มีการตีความอื่น วิธีการในขั้นนี้ ได้แก่ การรวบรวม และการนำเสนอ เป็นตาราง หรือ แผนภูมิ แล้วประกอบ การวิเคราะห์ด้วยสถิติระดับ สถิติพรรณนา(Descriptive Statistics) ซึ่งเป็นสถิติขั้นพื้นฐาน ได้แก่ ค่าเฉลี่ย (Mean) ร้อยละ (Percentage) ความถี่(Frequency) ส่วนเบี่ยงเบนมาตรฐาน(Standard Deviation) มัธยฐาน (Median) เปอร์เซ็นไทล์(Percentile) เป็นต้น ผลการวิเคราะห์ขั้นนี้เป็นการอธิบายว่าถึงปัจจุบันนี้สถานการณ์ของกิจการเป็นอย่างไร ไม่สามารถตีความนอกเหนือจากที่ปรากฏได้ ทั้งนี้หนังสือเล่มนี้ขอไม่กล่าวถึงเพราะขั้นนี้เป็นการจัดการ Data ขั้นต้น การทบทวนความรู้ส่วนนี้เป็นความรู้ที่เรียนผ่านมาแล้วในชั้นมัธยม ซึ่งปัจจุบันมีโปรแกรมคอมพิวเตอร์ประเภท Data Visualization หลากหลายเพื่อใช้สำหรับการวิเคราะห์ระดับ Descriptive analytics นี้

2) Diagnostic analytics เป็นการวิเคราะห์สาเหตุหรือข้อสงสัย ที่ทำให้ข้อมูลมีลักษณะตามที่ได้พบในระดับ Descriptive analytics เรียกว่าเกิดมี ข้อสันนิษฐาน ภาษาสถิติเรียกว่า สมมติฐาน (Hypothesis) เช่น

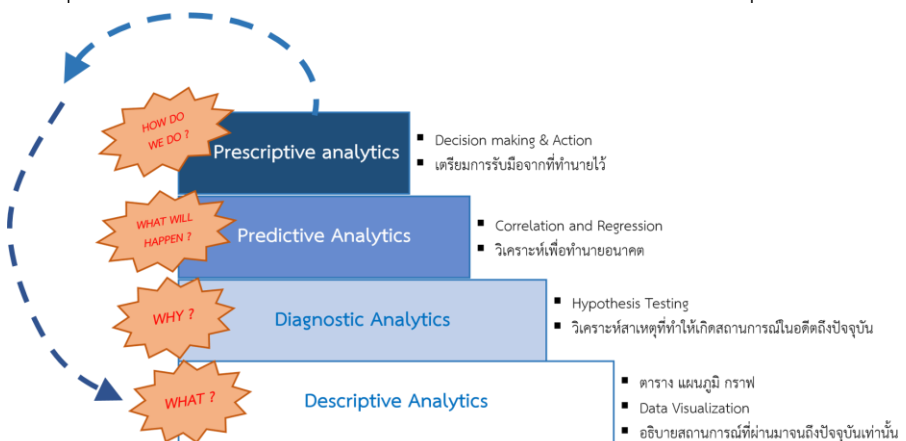
หลังจากตรวจสอบตารางแสดงยอดขายเบียร์ยี่ห้อหนึ่งในช่วงเดือนธันวาคม เกิดข้อสันนิษฐานว่า เดือนที่อากาศเย็นมีผลทำให้ยอดขายมากขึ้นกว่าเดือนที่อากาศร้อน ทำให้ต้องใช้วิธีการทางสถิติ คือ การทดสอบสมมติฐาน (Hypothesis Testing) เพื่อสนับสนุนข้อสันนิษฐานนี้ ซึ่งในวิธีการของการทดสอบสมมติฐานมีสถิติที่เกี่ยวข้อง หลายค่า ตามลักษณะของสมมติฐาน และลักษณะของ Data นั้นๆ เช่น ค่าสถิติ F , ค่าสถิติ t เป็นต้น การทดสอบสมมติฐาน เป็นสถิติระดับสถิติอนุมาน (Inferential Statistics) ซึ่งสถิตินี้จำแนกได้ เป็น 2 ประเภท คือ ประเภทที่ 1 สถิติแบบพารามेटริกซ์ (Parametric Statistics) คือ สถิติที่ให้ความสำคัญกับประชากรและการสุ่มตัวอย่าง ส่วนใหญ่กลุ่มตัวอย่างจะมีการแจกแจงปกติ และมีจำนวนมากพอ และ ประเภทที่ 2 สถิติแบบไม่ใช้พารามेटริกซ์ (Nonparametric Statistics) คือ สถิติที่ไม่เน้นลักษณะประชากรและกลุ่มตัวอย่างมีขนาดเล็ก ตัวแปรมีมาตรวัดระดับ Ordinal หรือ Nominal ก็ได้ ในที่นี้จะอธิบายเฉพาะสถิติอนุมาน (Inferential Statistics)แบบพารามेटริกซ์ (Parametric Statistics) เพราะเป็นพื้นฐานในการเข้าใจ Hypothesis Testing

3) Predictive analytics เป็นการวิเคราะห์เพื่อทำนายอนาคต เกิดจากการศึกษาความสัมพันธ์ของ Data จนสามารถทำนายเหตุการณ์ที่จะเกิดในอนาคต เช่น ศึกษาพบว่ายอดขายเบียร์มีความสัมพันธ์ทางบวกกับอัตราการว่างงาน และมีรายงานจากหน่วยงานที่เชื่อถือได้ว่าปีหน้าอัตราการว่างงานจะสูงขึ้น จึงทำนายยอดขายเบียร์ปีหน้าไว้ล่วงหน้าได้ เป็นต้น วิธีการของสถิติวิเคราะห์เบื้องต้น ในการศึกษาความสัมพันธ์ และทำนายข้อมูลใน

อนาคต หรือการพยากรณ์ (Prediction) คือ การวิเคราะห์สหสัมพันธ์และการถดถอย (Correlation and Regression Analysis)

4) Prescriptive analytics เป็นขั้นการวิเคราะห์สำหรับการเตรียมการรับมือกับสิ่งที่ได้จากขั้นการทำนาย เช่น ผลการทำนายยอดขายเบียร์ของเดือนธันวาคมปีหน้า พบว่า จะลดลง 50 % ดังนั้นการเตรียมการรับมือต่อยอดขายที่ลดลง ได้แก่การจัดโปรโมชั่นแบบต่างๆ เป็นต้น

ซึ่งจะพบว่า เมื่อถึงขั้นของ Prescriptive analytics แล้ว เช่น การจัดโปรโมชั่น แล้ว ก็ทำให้เกิด Data ขึ้นมาเพิ่มอีกจากกิจกรรมนี้ ย่อมเกิดการเปรียบเทียบ นำไปสู่การทดสอบข้อสันนิษฐานอื่นๆ เกิดกระบวนการทดสอบสมมติฐาน และทำนาย ขึ้นอีกไม่สิ้นสุด แสดงให้เห็นว่าในทุกๆกิจกรรมมีกระบวนการของ Data Analytics ตลอดเวลา ไม่สามารถหยุดนิ่งได้ กิจกรรมที่ไม่มีกระบวนการของ Data Analytics หรือหยุดที่ขั้นใดๆ ย่อมหมายถึงหยุดหรือยุติ หรือไม่มีการพัฒนา ย่อมไม่เท่าทันต่อการแข่งขันในโลกปัจจุบัน



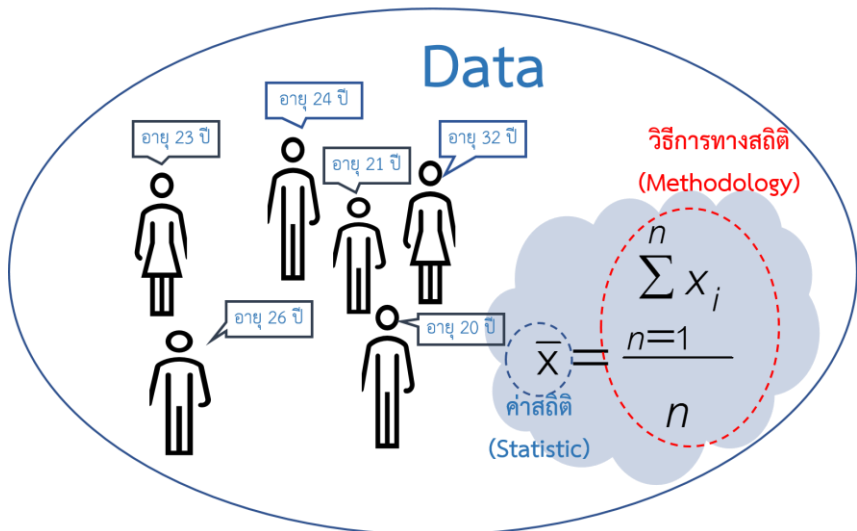
รูปที่ 1.1 กระบวนการของ Data Analytics

1.1 ความหมายและประเภทของสถิติ

สถิติ(Statistics) หมายถึง ศาสตร์หนึ่งซึ่งเป็นวิธีการทั้งหลายที่เกี่ยวข้องกับตัวเลขของกิจการหรือเหตุการณ์ใด ๆ ที่มีจุดมุ่งหมายให้ได้ผลสรุปเพื่อนำไปประกอบการตัดสินใจ (Decision Making) ในการดำเนินกิจการหรือเหตุการณ์นั้น ๆ ต่อไป สถิติจึงประกอบด้วยวิธีการทั้งหมดที่จะให้ได้ผลสรุปของกิจการหรือเหตุการณ์ ประกอบด้วย วิธีการเก็บรวบรวม Data (Collecting) การนำเสนอ (Presenting) การวิเคราะห์ และสรุปผล (Analysis & Conclusion)

ส่วนคำว่า “ค่าสถิติ (Statistic)” ซึ่งหมายถึง ค่าตัวเลข ที่ได้มาจากวิธีการทางสถิติ ด้วยวิธีใดวิธีหนึ่งโดยเฉพาะ เช่น ในการศึกษาอายุเฉลี่ยของลูกค้ากลุ่มหนึ่งที่นิยมซื้อสินค้าออนไลน์ชนิดหนึ่ง ได้อายุเฉลี่ยของลูกค้ากลุ่มเท่ากับ 25 ปี แสดงว่าในการศึกษานี้ได้ใช้วิธีการทางสถิติรวบรวม Data อายุของลูกค้าที่เป็นกลุ่มตัวอย่าง แล้วนำมาหาค่าเฉลี่ยของกลุ่มตัวอย่าง (Sample Mean; $\bar{x}$) ดังนั้น ค่าสถิติ (Statistic) ในการศึกษาครั้งนี้ คือค่าเฉลี่ย ($\bar{x}$) เท่ากับ 25 ปี ส่วนวิธีการทางสถิติ คือวิธีการคำนวณหาค่าเฉลี่ย

สรุปรวมความหมายของ สถิติวิเคราะห์(Statistical Analysis) หมายถึง วิธีการที่ประกอบด้วย ข้อมูล(Data) ค่าสถิติ(Statistic) และวิธีการทางสถิติ(Statistical Methodology) ที่เหมาะสมตามทฤษฎีสถิติ เพื่อให้ได้ข้อสรุป(Conclusion)ไปใช้ในการตัดสินใจ(Decision Making) ในกิจการนั้นๆ



รูปที่ 1.2 ส่วนประกอบของสถิติวิเคราะห์(Statistical Analysis)

สถิติแบ่งตามวิธีการได้ 2 ประเภท คือ

1) สถิติพรรณนา (Descriptive Statistics) หมายถึง วิธีการทั้งหลาย ในการอธิบาย และนำเสนอ Data ซึ่งปัจจุบันมีปริมาณมาก(Big Data) และไม่มีโครงสร้างที่ชัดเจน เช่น ข้อมูลจากสื่อสังคมออนไลน์ ข้อมูลส่วนบุคคล ข้อมูลลูกค้า ฯลฯ การนำมาจัดการให้เป็นระเบียบ เป็นหมวดหมู่ สวยงามชัดเจนเข้าใจง่าย ได้แก่ การสร้างตารางData การแจกแจงความถี่ (Frequency Distribution) การแสดงเป็นแผนภูมิ กราฟ(Chart , Graph) การหาค่าร้อยละ(Percentage) ค่าสูงสุด (Maximum) ค่าต่ำสุด (Minimum) ค่าเฉลี่ย (Mean) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ค่าสัดส่วน (Proportion) เป็นต้น เป็นวิธีการทางสถิติเบื้องต้น