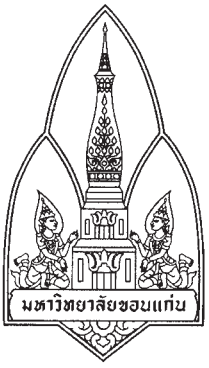
การวิเคราะห์ข้อมูลเชิงสำรวจ
Exploratory Data Analysis

รองศาสตราจารย์ ดร.สุพรรณี อึ้งปัญสัตวงศ์
สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น



การวิเคราะห์ข้อมูลเชิงสำรวจ

รองศาสตราจารย์สุพรรณี อึ้งปัญสัตวงศ์

คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

ได้รับทุนสนับสนุนการผลิตตำรา มหาวิทยาลัยขอนแก่น

ลำดับที่ 211

โดยมหาวิทยาลัยขอนแก่น พ.ศ.2566

ISBN (e-book) 978-616-438-815-4

สุพรรณี อึ้งปัญสัตวงศ์.
การวิเคราะห์ข้อมูลเชิงสำรวจ / สุพรรณี อึ้งปัญสัตวงศ์ -- พิมพ์ครั้งที่ 1. -- ขอนแก่น : คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 2566.
237 หน้า : ภาพประกอบ
1. สถิติวิเคราะห์. 2. ข้อมูล -- การวิเคราะห์ -- ระเบียบวิธีทางสถิติ. 3. คณิตศาสตร์สถิติ. 4. สถิติ.
(1) มหาวิทยาลัยขอนแก่น. คณะวิทยาศาสตร์. (2) ชื่อเรื่อง.
QA278 ส829 2566
ISBN (e-book) 978-616-438-815-4

จัดทำโดย ศูนย์นวัตกรรมการเรียนการสอน มหาวิทยาลัยขอนแก่น
ราคา 290 บาท
สงวนลิขสิทธิ์ โดยมหาวิทยาลัยขอนแก่น ตั้งแต่ปี พ.ศ. 2566

คำนำ

ตำราเล่มนี้ได้รวบรวมและจัดพิมพ์ขึ้นเป็นรูปเล่มที่สมบูรณ์ครั้งแรก ซึ่งปรับปรุงมาจากเอกสารประกอบการสอนในรายวิชา Sc613 207 การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis) อันเป็นวิชาที่เปิดสอนสำหรับ นักศึกษาในสาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น โดยการเรียบเรียงเนื้อหานี้ ได้ศึกษาจากตำราที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลเชิงสำรวจที่ได้รับการยอมรับและใช้กันอยู่ในปัจจุบัน ทั้งที่เป็นภาษาไทยและภาษาอังกฤษ วัตถุประสงค์ของการจัดพิมพ์ตำราเล่มนี้คือ เพื่อต้องการให้ผู้อ่านทราบถึงวิธีการต่างๆ ในการวิเคราะห์ข้อมูลด้วยเทคนิคง่ายๆ ซึ่งจะสามารถค้นหาคำตอบของปัญหาที่เกิดขึ้นในชุดข้อมูลเป็นการเบื้องต้นได้ โดยอาศัยวิธีการทางสถิติที่อาจไม่ต้องมีเงื่อนไขหรือข้อตกลงเบื้องต้นมากนัก เพื่อให้เกิดประโยชน์ในการใช้งานได้อย่างกว้างขวางต่อไป ในตำราชุดนี้ได้แสดงวิธีการคำนวณเพื่อวิเคราะห์ข้อมูลในการตอบปัญหาแต่ละเรื่องประกอบโดยละเอียด รวมทั้งได้มีการสรุปการแปลผลจากการวิเคราะห์ เพื่อใช้เป็นแนวทางสำหรับผู้สนใจที่จะใช้ความรู้เหล่านี้ ในการวิเคราะห์ข้อมูลต่อไป

ความสำเร็จของตำราเล่มนี้ เกิดขึ้นจากการได้รับทุนสนับสนุนการผลิตตำราใหม่ ซึ่งเป็นโครงการของมหาวิทยาลัยขอนแก่น พ.ศ. 2563 และการสนับสนุนสิ่งอำนวยความสะดวกต่างๆ ในการจัดพิมพ์ตำราเล่มนี้จากสาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น รวมถึง ดร.จิตรจิรา ไชยฤทธิ์ ที่ได้ช่วยตรวจทานและจัดทำรูปเล่มให้สมบูรณ์ยิ่งขึ้น ผู้เรียบเรียงขอขอบพระคุณมา ณ ที่นี้

รองศาสตราจารย์ ดร.สุพรรณี อึ้งปัญสัตวงศ์

มกราคม 2566

สารบัญ

	หน้า
บทนำ	1
บทที่ 1 ความรู้เบื้องต้นเกี่ยวกับการวิเคราะห์ข้อมูลเชิงสำรวจ	5
1. ความหมายของการวิเคราะห์ข้อมูลเชิงสำรวจ	5
2. กรอบของการวิเคราะห์ข้อมูลเชิงสำรวจ	5
3. วิธีการวิเคราะห์ข้อมูลเชิงสำรวจ	6
4. ประโยชน์ของการศึกษาการวิเคราะห์ข้อมูลเชิงสำรวจ	8
บทที่ 2 การนำเสนอข้อมูลด้วยแผนภาพลำต้นและใบ	9
1. แผนภาพลำต้นและใบ	9
2. การเปรียบเทียบชุดข้อมูล 2 กลุ่ม	12
3. กรณีข้อมูลซ้ำกันหลาย ๆ ค่า	15
4. การแสดงแผนภาพลำต้นและใบที่ค่าของข้อมูลดิบมีลักษณะเป็นจุดทศนิยม	16
5. ค่ามัธยฐานและตำแหน่งของมัธยฐาน	18
6. การนำเสนอข้อมูลที่เป็นบวกและลบ	24
7. การแสดงข้อมูลที่แตกต่างจากพวก	26
8. การออกแบบการนำเสนอข้อมูล	27
9. การออกแบบการนำเสนอข้อมูลแบบกราฟการกระจายในรูปแบบของแผนภาพลำต้น	28
แบบฝึกหัดท้ายบทที่ 2	31
บทที่ 3 การนำเสนอข้อมูลด้วยค่าตัวอักษร	35
1. มัธยฐาน ฮินเจส และค่าสรุปอื่น ๆ (Median Hinges and Other Summary Values)	35
2. ค่าตัวอักษร (Letter Value)	39
3. การแสดงผลของค่าตัวอักษร (Displaying the letter values)	42
แบบฝึกหัดท้ายบทที่ 3	46
บทที่ 4 แผนภาพกล่อง	47
1. ความหมายของแผนภาพกล่อง (Box-plot)	47
2. การวัดการกระจายของข้อมูล	48

สารบัญ (ต่อ)

	หน้า
3. วิธีการอ่านแผนภาพกล่อง	49
4. ขั้นตอนการสร้างแผนภาพกล่อง	50
5. การใช้แผนภาพกล่องเปรียบเทียบการแจกแจงของข้อมูลหลายชุด	52
แบบฝึกหัดท้ายบทที่ 4	55
บทที่ 5 ตารางเครื่องหมาย	63
1. การแสดงตาราง	65
2. ตารางเครื่องหมายและแผนภาพกล่อง	74
แบบฝึกหัดท้ายบทที่ 5	75
บทที่ 6 การแปลงข้อมูล	77
1. เหตุผลของการแปลงข้อมูล	78
2. กำลังของการแปลง (Power Transformations)	79
3. การแปลงข้อมูลเพื่อความสมมาตร (Transforming for Symmetry)	80
4. การแปลงข้อมูลเพื่อวัตถุประสงค์อื่นๆ	90
แบบฝึกหัดท้ายบทที่ 6	101
บทที่ 7 เส้นต้านทาน	103
1. การสร้างเส้นต้านทานจากข้อมูล 3 กลุ่ม	103
2. การสร้างเส้นต้านทานด้วยวิธีอื่น ๆ	123
2.1 วิธีการวิเคราะห์เส้นต้านทาน โดยวิธีธิลส์อินคอมพลีท (Theil's Incomplete Method)	123
2.2 วิธีการวิเคราะห์เส้นต้านทาน โดยวิธีของวาลด์ (Wald's Method)	126
2.3 วิธีของแนร์และสีวาสตาวา (Method of Nair and Shrivastava)	128
2.4 วิธีบาร์ทเล็ทส์ (Bartletts Method)	130
3. เส้นต้านทานกับการถดถอย ด้วยวิธีกำลังสองน้อยที่สุด	132
แบบฝึกหัดท้ายบทที่ 7	138

สารบัญ (ต่อ)

	หน้า
บทที่ 8 การวิเคราะห์ตารางสองทางด้วยค่ามัธยฐาน	141
1. รูปแบบของตารางสองทาง (Two –Way Tables)	141
2. การวิเคราะห์ด้วยวิธี Median Polish	143
3. การวิเคราะห์ด้วยวิธีตารางรวมสี่เหลี่ยม (Square Combining Table)	155
4. การวิเคราะห์โดยใช้ ค่าเฉลี่ย (Analysis by Means)	161
5. การเปรียบเทียบการวิเคราะห์ด้วยวิธีการขัดมัธยฐาน (Median Polish) กับการวิเคราะห์ความแปรปรวน (ANOVA)	168
แบบฝึกหัดท้ายบทที่ 8	169
บทที่ 9 การวิเคราะห์ตารางสามทางด้วยค่ามัธยฐาน	173
1. โครงสร้างของตารางสามทาง (Three –Way Tables)	173
2. ตัวแบบสำหรับการวิเคราะห์ในตารางสามทาง	174
3. การวิเคราะห์ด้วยวิธี Median Polish	174
4. วิธีการวิเคราะห์โดยใช้ค่าเฉลี่ย (Analysis using means)	183
แบบฝึกหัดท้ายบทที่ 9	189
บทที่ 10 การถดถอยแบบแกร่ง	191
10.1 บทนำ	191
10.2 ค่าผิดปกติในการวิเคราะห์การถดถอย	194
10.3 จุด Breakdown และ การประมาณค่าแบบแกร่ง	202
10.4 วิธีประมาณค่า แบบ Maximum Likelihood (ML)	207
แบบฝึกหัดท้ายบทที่ 10	224
บรรณานุกรม	227
ดัชนี	231

สารบัญแผนภาพ

		หน้า
แผนภาพที่ 2.1	การกำหนดเลขหน้าของแผนภาพลำต้นและใบ	10
แผนภาพที่ 2.2	แผนภาพลำต้นและใบ	10
แผนภาพที่ 2.3	แผนภาพลำต้นและใบที่จัดเรียงใหม่	11
แผนภาพที่ 2.4	แผนภาพลำต้นและใบ ข้อมูลส่วนสูง (ซม.) ของนักเรียนชั้นประถมปีที่ 6	12
แผนภาพที่ 2.5	แผนภาพลำต้นและใบ จากข้อมูลทั้ง 2 ชุด โดยเรียงลำดับค่าข้อมูลจากน้อยไปมาก	13
แผนภาพที่ 2.6	แสดงข้อมูลดิบและแผนภาพลำต้นและใบ ของตัวอย่างที่ 2.4	14
แผนภาพที่ 2.7	แผนภาพลำต้นและใบของคะแนนเพศชายและเพศหญิงหลังทำการทดลอง	16
แผนภาพที่ 2.8	แผนภาพลำต้นและใบแสดงการแจกแจงของข้อมูลในตัวอย่างที่ 2.6	17
แผนภาพที่ 2.9	ตัวอย่างของการแบ่งแผนภาพลำต้นและใบ ของชุดข้อมูลตัวอย่างที่ 2.7	18
แผนภาพที่ 2.10	แผนภาพลำต้นและใบ ของชุดข้อมูลมลพิษทางอากาศของก๊าซไฮโดรคาร์บอน จาก 60 เมืองของสหรัฐอเมริกา	20
แผนภาพที่ 2.11	แผนภาพลำต้นและใบของความเป็นกรดของ 26 ตัวอย่างของน้ำฝนที่เก็บได้จากสถานที่ใน Allegheny County, Pennsylvania ตั้งแต่เดือน ธันวาคม 1973 ถึง มิถุนายน 1974	21
แผนภาพที่ 2.12	แผนภาพลำต้นและใบของชุดข้อมูลในตัวอย่างที่ 2.9	23
แผนภาพที่ 2.13	แผนภาพลำต้นและใบของอุณหภูมิเฉลี่ยในหน่วย °C ของ 60 เมืองในสหรัฐอเมริกา ในเดือนมกราคม	25
แผนภาพที่ 2.14	แผนภาพลำต้นและใบของข้อมูลจากตัวอย่างที่ 2.10 ที่ไม่ต้องการจะแยกค่าที่สูงมากออกจากพวก	26
แผนภาพที่ 2.15	แผนภาพลำต้นและใบของข้อมูลในตัวอย่างที่ 2.12	28
แผนภาพที่ 2.16	แผนภาพลำต้นและใบของคะแนนคณิตศาสตร์ของนักเรียนจำนวน 41 คน	29
แผนภาพที่ 4.1	แผนภาพกล่องของชุดข้อมูลที่มีลักษณะสมมาตร	47
แผนภาพที่ 4.2	แผนภาพกล่องของชุดข้อมูลที่มีลักษณะเบ้ขวา	47
แผนภาพที่ 4.3	แผนภาพกล่องของชุดข้อมูลที่มีลักษณะเบ้ซ้าย	48
แผนภาพที่ 4.4	แผนภาพกล่องของชุดข้อมูลที่มีลักษณะกระจายมากและเบ้ขวา	48
แผนภาพที่ 4.5	แผนภาพกล่องของชุดข้อมูลที่มีลักษณะกระจายน้อยและเบ้ซ้าย	48

สารบัญแผนภาพ (ต่อ)

		หน้า
แผนภาพที่ 4.6	แผนภาพกล่องของชุดข้อมูลทั่วๆ ไป	49
แผนภาพที่ 4.7	รูปแบบของแผนภาพกล่องของชุดข้อมูล	50
แผนภาพที่ 4.8	ส่วนประกอบของแผนภาพกล่องของชุดข้อมูล	51
แผนภาพที่ 4.9	แผนภาพกล่องของคะแนนสอบของนักศึกษา 17 คน	52
แผนภาพที่ 4.10	แผนภาพกล่องของความต้านทานแรงดึงของเส้นใยสังเคราะห์	54
แผนภาพที่ 5.1	ขอบเขตสำหรับการกำหนดเครื่องหมาย (สัญลักษณ์)	66
แผนภาพที่ 5.2	ขอบเขตการกำหนดเครื่องหมายสำหรับข้อมูลในตาราง 5.1	67
แผนภาพที่ 5.3	ขอบเขตการกำหนดเครื่องหมายสำหรับข้อมูลในตัวอย่างที่ 5.3	73
แผนภาพที่ 6.1	แผนภาพลำต้นและใบ (stem-and-leaf display) ของข้อมูลจากตาราง 6.2	82
แผนภาพที่ 6.2	ฟังก์ชันเส้นโค้งที่มีลักษณะเป็นตัวแบบเชิงเส้นแท้จริง (Intrinsically linear model)	91
แผนภาพที่ 6.3	แผนภาพกระจายตัวแปร X กับ ตัวแปร Y	92
แผนภาพที่ 6.4	แผนภาพกระจายระหว่างค่าพยากรณ์กับความคลาดเคลื่อน และแผนภาพกระจายระหว่างตัวแปรอิสระกับความคลาดเคลื่อน	93
แผนภาพที่ 6.5	แผนภาพกระจายระหว่างตัวแปรทั้งสองหลังจากการแปลงด้วยลอการิทึม	94
แผนภาพที่ 6.6	แผนภาพการกระจาย	96
แผนภาพที่ 6.7	แผนภาพการกระจายระหว่าง e_i , $\hat{y}_{i1}$	98
แผนภาพที่ 6.8	แผนภาพการกระจายระหว่าง e_i , x_i	99
แผนภาพที่ 6.9	แผนภาพการกระจายระหว่าง e_i , $\hat{y}_{i2}$	99
แผนภาพที่ 6.10	แผนภาพการกระจายระหว่าง y กับ 1/x	100
แผนภาพที่ 6.11	แผนภาพการกระจายระหว่าง e_i กับ 1/x	100
แผนภาพที่ 7.1	ลักษณะการแบ่งกลุ่มข้อมูลออกเป็น 3 กลุ่ม	104
แผนภาพที่ 7.2	แผนภาพการกระจายระหว่างอุณหภูมิเฉลี่ยประจำปีกับดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิง ในปี 1965	107
แผนภาพที่ 7.3	แผนภาพการกระจายแสดงค่าใช้จ่ายกับยอดขาย	135
แผนภาพที่ 7.4	แผนภาพแสดงการเปรียบเทียบยอดขายและค่าพยากรณ์	137

สารบัญแผนภาพ (ต่อ)

หน้า

แผนภาพที่ 8.1 แผนภาพลำต้นและใบของค่าความคลาดเคลื่อนจากการทำการวิเคราะห์ซ้ำด้วยการขัดมัธยฐานของข้อมูลอัตราการตายของเด็กทารกในสหรัฐอเมริกา ปี ค.ศ. 1964 – 1966 154

แผนภาพที่ 8.2 แผนภาพลำต้นและใบของค่าความคลาดเคลื่อนจากการวิเคราะห์ในตาราง 8.17 ของข้อมูลอัตราการตายของเด็กทารกในสหรัฐอเมริกา ปี ค.ศ. 1964 – 1966 161

แผนภาพที่ 8.3 เปรียบเทียบค่าความคลาดเคลื่อนจากการวิเคราะห์โดยใช้ค่าเฉลี่ยและวิเคราะห์โดยใช้ Median Polishของภาวะอัตราการตายของเด็กทารก (บันทึกหลัง “hi”11.2 เป็นตัวเดียวที่มีความคลาดเคลื่อนมาก 11.2 เป็นค่าที่มาก สามารถแสดงได้สะดวกในแผนภาพต้นและใบ) ของข้อมูลอัตราการตายของเด็กทารกในสหรัฐอเมริกา ปี ค.ศ. 1964 – 1966 167

แผนภาพที่ 10.1 แผนภาพการกระจายกรณีข้อมูลมีค่าผิดปกติ 192

แผนภาพที่ 10.2 การเรียงตัวของจุด 5 จุด บนเส้นถดถอยกำลังสองน้อยที่สุดและแสดงค่าผิดปกติในแกน Y 197

แผนภาพที่ 10.3 การเรียงตัวของจุด 5 จุด บนเส้นถดถอยกำลังสองน้อยที่สุดและแสดงค่าผิดปกติในแกน X (จุด leverage) 198

แผนภาพที่ 10.4 ลักษณะจุด Leverage ที่ดี 200

แผนภาพที่ 10.5 อธิบายการชุดข้อมูลที่ลงจุดตัวแปร (x_{i1}, y_{i2}) ของการถดถอยมี 2 จุด (จุดที่ระบุโดยวงกลมประ) 201

แผนภาพที่ 10.6 แสดงความแข็งแรงของ การถดถอย L_1 เทียบกับค่าผิดปกติในแกน Y และ (b) แสดงความไวของการถดถอย L_1 กับค่าผิดปกติในแกน X 204

แผนภาพที่ 10.7 (a) แสดงความทนทาน (Robustness) ของการถดถอยแบบ LMS เทียบกับ ค่าผิดปกติในแกน y และ รูป (b) แสดงค่าผิดปกติในแกน x 211

แผนภาพที่ 10.8 ข้อมูลการทดลองของพืชที่ใช้วิธี LS (เส้นประ) และ LMS (เส้นทึบ) 216

แผนภาพที่ 10.9 จากข้อมูลในตาราง 10.1 แต่มีค่าผิดปกติ 1 ค่า เส้นประคือวิธีกำลังสองน้อยที่สุด (LS) ที่เหมาะสม และเส้นทึบคือวิธี LMS 218

สารบัญแผนภาพ (ต่อ)

หน้า

แผนภาพที่ 10.10 จำนวนการโทรศัพท์ระหว่างประเทศ ของประเทศเบลเยียม ในปี 1950-1973 แสดงโดยวิธี LS (เส้นประ) และวิธี LMS ที่เหมาะสม (เส้นทึบ) 221

สารบัญตาราง

		หน้า
ตาราง 1.1	ตารางแจกแจงความถี่	7
ตาราง 5.1	อัตราการตายของผู้ชายด้วยสาเหตุต่าง ๆ จากการสูบบุหรี่ ต่อ 1000 คน แบ่งเป็นชนิดของปริมาณการสูบบุหรี่	64
ตาราง 5.2	เป็นผลลัพธ์ที่ได้จากการคำนวณ จะสามารถกำหนดเป็นตารางเครื่องหมาย	68
ตาราง 5.3	ระยะเวลาที่รอดชีวิตของสัตว์หลังที่ได้รับสารพิษ ทั้ง 3 ชนิด และ ได้รับการรักษาจากสิ่งทดลอง (treatment)	70
ตาราง 5.4	ค่าต่ำสุดในแต่ละเซลล์	71
ตาราง 5.5	ค่าสูงสุดในแต่ละเซลล์	71
ตาราง 5.6	ข้อมูลจำนวนการเกิดของเด็กทารกในภาคเหนือของประเทศไทย พ.ศ. 2544 – 2553	71
ตาราง 5.7	ตารางเครื่องหมายของการเกิดของเด็กทารกในภาคเหนือของประเทศไทย พ.ศ. 2544 – 2553	73
ตาราง 6.1	ค่ายกกำลัง (Power) และฟังก์ชันการแปลงข้อมูล	80
ตาราง 6.2	ปริมาณน้ำฝนที่วัดได้ในช่วงเดือนมีนาคมในปี 2530	82
ตาราง 6.3	ชุดค่าตัวอักษร (letter value display) ของปริมาณน้ำฝนที่วัดได้ ในช่วงเดือนมีนาคม พ.ศ. 2530	83
ตาราง 6.4	ข้อมูลดิบซึ่งเรียงลำดับจากน้อยไปหามาก และนำข้อมูลดิบมาทำการแปลงข้อมูล โดยใช้ฟังก์ชันรากที่สอง (Square-root function) ฟังก์ชันล็อก (log function) และฟังก์ชันเศษส่วนกลับ (Reciprocal function)	85
ตาราง 6.5	ชุดค่าตัวอักษรของข้อมูลปริมาณน้ำฝนที่วัดได้ในช่วงเดือนมีนาคม พ.ศ. 2530 สำหรับการแปลงค่า 3 รูปแบบ (ฟังก์ชัน)	86
ตาราง 6.6	สรุปค่ากลาง (ค่า Mid) ของการใช้ฟังก์ชันการแปลงค่ารูปแบบต่าง ๆ ของข้อมูลปริมาณน้ำฝนที่วัดได้ในช่วงเดือนมีนาคม พ.ศ. 2530	88
ตาราง 6.7	การกระจายสำหรับการแจกแจงแบบปรกติมาตรฐาน (Standard Gaussian distribution)	89
ตาราง 6.8	ฟังก์ชันเชิงเส้นโค้งและฟังก์ชันเชิงเส้นตรงที่แปลงรูป	91
ตาราง 6.9	ค่าสังเกตจากการทดลองปั่นกระแสไฟฟ้าด้วยกังหันลม	95

สารบัญตาราง (ต่อ)

หน้า

ตาราง 6.10 เปรียบเทียบค่าทำนายและค่าความคลาดเคลื่อนจากตัวแบบเชิงเส้นตรง และตัวแบบที่แปลงรูป 97

ตาราง 7.1 อุณหภูมิเฉลี่ยประจำปีกับดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านม ในเพศหญิง ในปี 1965 105

ตาราง 7.2 อุณหภูมิเฉลี่ยประจำปีกับดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านม ในเพศหญิง ในปี 1965 ที่ได้เรียงลำดับแล้วและทำการแบ่งชุดข้อมูล ออกเป็น 3 กลุ่ม 106

ตาราง 7.3 ผลการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิง เมื่อแทนค่าอุณหภูมิเฉลี่ยประจำปีของแต่ละหน่วยตัวอย่าง รอบที่ 1 108

ตาราง 7.4 ค่าความคลาดเคลื่อนรอบที่ 1 (First Residual) จากสมการทำนายดัชนี การเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิงของแต่ละหน่วยตัวอย่าง 109

ตาราง 7.5 ผลการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิง เมื่อแทนค่าอุณหภูมิเฉลี่ยประจำปีของแต่ละหน่วยตัวอย่าง รอบที่ 2 111

ตาราง 7.6 ค่าความคลาดเคลื่อนรอบที่ 2 (Second Residual) จากสมการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิงของแต่ละหน่วยตัวอย่าง 112

ตาราง 7.7 ผลการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิง เมื่อแทนค่าอุณหภูมิเฉลี่ยประจำปีของแต่ละหน่วยตัวอย่าง รอบที่ 3 113

ตาราง 7.8 ค่าความคลาดเคลื่อนรอบที่ 3 (Third Residual) จากสมการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิงของแต่ละหน่วยตัวอย่าง 114

ตาราง 7.9 ผลการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิง เมื่อแทนค่าอุณหภูมิเฉลี่ยประจำปีของแต่ละหน่วยตัวอย่าง รอบที่ 4 116

ตาราง 7.10 ค่าความคลาดเคลื่อนรอบที่ 4 (Forth Residual) จากสมการทำนาย ดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิงของแต่ละหน่วยตัวอย่าง 117

ตาราง 7.11 ค่าความคลาดเคลื่อนรอบสุดท้าย จากสมการทำนายดัชนีการเสียชีวิต ของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิงของแต่ละหน่วยตัวอย่าง 119

ตาราง 7.12 ผลการทำนายค่าดัชนีการเสียชีวิตของผู้ป่วยโรคมะเร็งเต้านมในเพศหญิง ของแต่ละหน่วยตัวอย่าง ที่เหมาะสมที่สุด 120

สารบัญตาราง (ต่อ)

		หน้า
ตาราง 7.13	ตารางสรุปผลการวิเคราะห์ในภาพรวม ของการกระทำซ้ำทั้งหมด 4 รอบ	122
ตาราง 7.14	ยอดขายเทียบกับค่าใช้จ่ายในการโฆษณาของบริษัทวันเฉลิม จำกัด	135
ตาราง 7.15	การวิเคราะห์ข้อมูลเพื่อหาสมการเส้นต้านทาน	136
ตาราง 8.1	รูปแบบและสัญลักษณ์ที่ใช้ในตารางสองทาง	141
ตาราง 8.2	ตาราง ANOVA ของตัวแบบเชิงบวก	142
ตาราง 8.3	ตารางการวิเคราะห์การขัดมัธยฐานทางแถว ณ รอบที่ n (Row median polish at iteration n)	146
ตาราง 8.4	ตารางการวิเคราะห์การขัดมัธยฐานทางสดมภ์ ณ รอบที่ n (Column median polish at iteration n)	146
ตาราง 8.5	อัตราการตายของเด็กทารกในสหรัฐอเมริกา ปี ค.ศ. 1964 – 1966 โดยจำแนกตามภูมิภาคมี 4 ระดับ และระดับการการศึกษาของบิดา มี 5 ระดับ (บันทึกจำนวนการตายต่อการเกิดรอด 1000 ราย)	147
ตาราง 8.6	อัตราการตายของเด็กทารก และค่ามัธยฐานในแนวแถว ซึ่งพร้อมที่จะทำการขัดมัธยฐานซ้ำในครั้งแรก (The fist iteration of median polish n=1)	148
ตาราง 8.7	อัตราการตายของเด็กทารกหลังจากการทำ polish ครั้งแรกในแนวแถว (n=1)	149
ตาราง 8.8	อัตราการตายของเด็กทารกหลังจากการทำ polish ครั้งแรกในแนวสดมภ์ (n=1)	150
ตาราง 8.9	อัตราการตายของเด็กทารกหลังจากการทำ median polish ครั้งที่ 2 (n=2) ในแนวแถว	151
ตาราง 8.10	อัตราการตายของเด็กทารกหลังจากการทำ polish ครั้งที่ 2 ในแนวสดมภ์	152
ตาราง 8.11	อัตราการตายของเด็กทารกหลังจากการทำ polish ครั้งที่ 3 (n=3) ในแนวแถว	154
ตาราง 8.12	ผลต่างของแต่ละคู่ของสดมภ์ (Column differences) จากข้อมูลในตาราง 8.5	156
ตาราง 8.13	ตารางรวมสี่เหลี่ยมสำหรับสดมภ์จากข้อมูลในตาราง 8.5 (Square Combing Table for the columns)	156
ตาราง 8.14	ผลต่างของแต่ละคู่ของแถว (Row differences) จากข้อมูลในตาราง 8.5	157
ตาราง 8.15	ตารางรวมสี่เหลี่ยมสำหรับแถวจากข้อมูลในตาราง 8.5 (Square combing table for the row)	158
ตาราง 8.16	ชุดข้อมูลในตาราง 8.5 ที่ได้หักผลกระทบเนื่องจากอิทธิพลของแถวและสดมภ์ออกไปแล้ว	159

สารบัญตาราง (ต่อ)

หน้า

ตาราง 8.17 ตารางรวมสี่เหลี่ยมของความคลาดเคลื่อน สำหรับข้อมูลตาราง 8.5 (Square-combining – table fit with residuals) 160

ตาราง 8.18 ตารางผลต่างของแต่ละค่าข้อมูลในตาราง 8.5 กับค่าเฉลี่ย (22.82) และเราจะสามารถหาค่าเฉลี่ยของแต่ละสดมภ์ (ได้ $_{i}$; i = 1, 2,..., r = 4) 164

ตาราง 8.19 ผลต่างของแต่ละค่าในแต่ละแถวกับค่าเฉลี่ยของแถวนั้นๆ (ทั้งหมด 4 แถว) 165

ตาราง 8.20 ค่าความคลาดเคลื่อนของแต่ละเซลล์ ของข้อมูลในตาราง 8.5 หลังจากการวิเคราะห์บนค่าเฉลี่ย 166

ตาราง 9.1 ข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริม 175

ตาราง 9.2 ข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริมและ column medians จากการหา median ในแต่ละระดับของไลซีน 176

ตาราง 9.3 ข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริมหลังจากขจัดค่ามัธยฐานในแต่ละระดับของไลซีน 178

ตาราง 9.4 จัดตาราง 9.3 ใหม่ โดยสร้าง Column ของเมไทโอนีนใหม่และหา Column median ซึ่งเกิดจากข้อมูลใน Column ที่สร้างใหม่ 178

ตาราง 9.5 ข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริมหลังจากขจัดค่ามัธยฐานในแต่ละระดับของเมไทโอนีน 178

ตาราง 9.6 จัดตาราง 9.5 ใหม่ โดยสร้าง Column ของโปรตีนใหม่และหา Column median ในระดับของโปรตีน 179

ตาราง 9.7 ข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริมหลังจากนำค่าสังเกตลบออกจากค่ามัธยฐาน ในตาราง 9.6 180

ตาราง 9.8 แสดงค่า Main-effects ของข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริมหลังจากทำ 1 รอบ 180

ตาราง 9.9 ผลขั้นสุดท้ายของการวิเคราะห์ผลกระทบข้อมูลอัตราการเติบโตของสุกรจากการให้อาหารเสริมโดยวิธี Median Polish 181

ตาราง 9.10 แผนภาพลำต้นและใบแสดงค่า residuals หลังการทำซ้ำครั้งสุดท้ายของการวิเคราะห์ผลกระทบหลักของข้อมูลน้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรโดย Median Polish (median at 0, fourths at -9 and 14,25) 181

สารบัญตาราง (ต่อ)

หน้า

ตาราง 9.11 ผลการเปลี่ยนแปลงของผลลัพธ์ในแต่ละรอบ ในกรณีวิเคราะห์ Median – Polish ที่มีผลกระทบหลักอย่างเดียวของข้อมูลการให้อาหารสุกร 182

ตาราง 9.12 ข้อมูลน้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรและ column means 184

ตาราง 9.13 ข้อมูลน้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรและอิทธิพลในแต่ละระดับของไลซีน 184

ตาราง 9.14 จัดตาราง 9.13 ใหม่ โดยสร้าง Column ของ เมไทโอนีน ใหม่และหา Column means ซึ่งเกิดจากข้อมูลใน Column ที่สร้างใหม่ 185

ตาราง 9.15 ข้อมูลน้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรและอิทธิพลในแต่ละระดับของ เมไทโอนีน 185

ตาราง 9.16 จัดตาราง 9.14 ใหม่ โดยสร้าง Column ของโปรตีนใหม่และหา Column means ในระดับของโปรตีน 186

ตาราง 9.17 ข้อมูลน้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรและอิทธิพลในแต่ละระดับของโปรตีน 187

ตาราง 9.18 ค่า residuals r_{ijk} และอิทธิพลในแต่ละระดับของ ไลซีน, เมไทโอนีนและโปรตีน 188

ตาราง 10.1 ชุดข้อมูลของพืชแปลงทดลอง 214

ตาราง 10.2 ตารางแสดงการคำนวณโดยวิธีกำลังสองน้อยที่สุด 215

ตาราง 10.3 ตารางแสดงการคำนวณโดยวิธีกำลังสองน้อยที่สุด 217

ตาราง 10.4 จำนวนการโทรศัพท์ระหว่างประเทศของเบลเยียม 220

ตาราง 10.5 แสดงข้อมูล และการคำนวณหาตัวประมาณค่าของสมการถดถอย โดยวิธีกำลังสองน้อยที่สุด 221

บทนำ

โดยทั่วไป ในวงวิชาการทางสถิติ เรามักจะแบ่งกลุ่มของการวิเคราะห์ข้อมูลทางสถิติออกเป็น 3 กลุ่ม ได้แก่

1. การวิเคราะห์ข้อมูลเชิงสำรวจ เป็นวิธีการทางสถิติที่พัฒนาโดยกลุ่มนักคณิตศาสตร์ที่พยายามคิดหาวิธีที่จะนำมาใช้ในการวิเคราะห์ข้อมูลทางสถิติ โดยไม่ต้องคำนึงถึงข้อตกลงเบื้องต้น เช่น เรื่องของรูปแบบการแจกแจงของข้อมูล เป็นต้น ซึ่งหมายถึงการนำเอาข้อมูลที่มีนั้นมาอธิบายโดยอาศัยเทคนิคต่างๆ ได้แก่

1. การทำให้เห็นเด่นชัด (Display)
2. ความคลาดเคลื่อน หรือ ส่วนเศษเหลือ (Residual)
3. การแสดงข้อมูลอีกครั้งในรูปแบบใหม่ (Re-expression)
4. ความต้านทาน (Resistance)

2. การอนุมานเชิงสถิติ (Inferential Statistics) และการวิเคราะห์นอนพาราเมตริก (Nonparametric Statistics)

1. การอนุมานเชิงสถิติ (Inferential Statistics) เป็นกลุ่มของการวิเคราะห์ทางสถิติ ซึ่งหลักการคือ จะมองค่าพารามิเตอร์ (parameter) ของชุดข้อมูลหรือของตัวแปรที่สนใจศึกษาเป็นค่าคงที่ที่ไม่ทราบค่า

2. การวิเคราะห์นอนพาราเมตริก (Non Parametric) เป็นการวิเคราะห์ทางสถิติที่ไม่สนใจค่าของ พารามิเตอร์ (parameter) ของข้อมูลหรือของตัวแปรที่สนใจศึกษา ที่เรียกว่าการแจกแจงอิสระ (Distribution Free)

โดยที่ใน 2 เรื่องนี้เราจะต้องเชื่อว่าข้อมูลของตัวแปรที่เรามีอยู่นี้ต้องได้มาโดยการสุ่มหรือต้องเป็น random variable (Johnson, R. A. And Wichern, D. W. 2002) ดังนั้น ข้อมูลที่จะนำมาวิเคราะห์นี้ จึงต้องเป็นข้อมูลเชิงสุ่ม นั่นคือ ทั้ง 2 เรื่องข้างต้น จะวิเคราะห์ได้ก็ต้องเชื่อได้ว่าตัวอย่างที่เรามีอยู่นั้นต้องเป็นตัวอย่างสุ่ม (Random Sample) เพียงแต่ เราต้องมั่นใจว่าข้อมูลเหล่านั้น จะต้องเป็นตัวแทนของประชากรที่เราต้องการศึกษา ดังนั้นในส่วนของการสุ่มตัวอย่างซึ่งเป็นส่วนที่จะทำให้ได้มาซึ่งข้อมูลที่จะนำไปวิเคราะห์ เราจะทำโดยวิธีใดก็ได้ขอเพียงให้เป็นการสุ่มภายใต้ภาวะน่าจะเป็นก็เพียงพอแล้ว

ดังนั้นหากข้อมูลที่จะนำมาใช้ในการวิเคราะห์นี้ได้มาโดยวิธีการสุ่มตัวอย่างแบบบังเอิญ แบบเจาะจง แบบโควต้า แบบลูกบอลหิมะหรือแบบที่เป็นเทคนิคในกลุ่มที่ไม่ใช้ความน่าจะเป็น เราก็จะสามารถเลี่ยงมาเลือกใช้วิธีการวิเคราะห์ข้อมูลในกลุ่ม การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis) นี้ได้เลย

3. การอนุมานแบบเบส์ (Bayesian Inference) เป็นการวิเคราะห์เชิงสถิติที่เริ่มมีบทบาทมากขึ้นในปัจจุบัน โดยมีแนวคิดเกี่ยวกับพารามิเตอร์ที่แตกต่างไปจากการอนุมานเชิงสถิติแบบทั่วไปที่เรารู้จัก กล่าวคือ ในการวิเคราะห์นี้ เราจะมองว่า พารามิเตอร์ของประชากรนั้นเป็นตัวแปรสุ่ม (Random Variable) นั่นเอง

ประวัติความเป็นมาของ EDA

เมื่อปี ค.ศ. 1962 John W. Tukey นักคณิตศาสตร์จาก Bell Labs (Lab ที่สร้างนวัตกรรมทางการสื่อสาร คอมพิวเตอร์ที่เราใช้กันในปัจจุบัน) เขียนบทความที่มีอิทธิพลต่อโลกของข้อมูลชื่อ The Future of Data Analysis ที่ชี้ว่าเครื่องมือทางสถิติในยุคนั้นมีความสามารถไม่เพียงพอกับการวิเคราะห์ข้อมูลที่มีอยู่อย่างมากมาย คำว่า Data Analysis เป็นคำใหม่ที่ไม่ใช่แค่เป็นการหาข้อสรุป (Inference) โดยเทคนิคทางสถิติเพียงอย่างเดียว แต่รวมไปถึงการเก็บข้อมูล การเตรียมและการแปลความหรือการแปลผล (ซึ่งถ้าสังเกตดูก็เป็นกระบวนการเดียวกันกับ Data Science ในยุคปัจจุบันนั่นเอง)

นอกจากนี้ Tukey ยังพูดถึงการสร้างความคิดใหม่ในการสอน การวิเคราะห์มีการตั้งคำถามใหม่ ๆ การลองเปลี่ยนข้อจำกัดของกฎเกณฑ์เก่าหรือการทำให้หลุดพ้นจากกรอบความคิดเก่าไปสู่ความคิดในรูปแบบใหม่ ๆ หลังจากนั้นในปี ค.ศ. 1977 Tukey ได้ตีพิมพ์หนังสือที่มีอิทธิพลกับงานด้านข้อมูลชื่อ “Exploratory Data Analysis” ที่เน้นการใช้ข้อมูลมาช่วยในการตั้งสมมติฐาน

ต่อมาในช่วงปี ค.ศ. 1970 - 1990 ระบบฐานข้อมูล (Database) เข้ามามีบทบาทมากขึ้น พร้อมทั้งมีการนำคอมพิวเตอร์มาช่วยในการวิเคราะห์ข้อมูลจึงเป็นยุคที่เริ่มได้ใช้คำว่า “เหมืองข้อมูล” (Data Mining) อย่างแพร่หลาย โดยในปี ค.ศ. 1996 Usama Fayyad, Gregory Piatetsky-Shapiro, และ Padhraic Smyth ได้ตีพิมพ์บทความชื่อว่า “From Data Mining to Knowledge Discovery in Databases.” ที่เล่าถึงกระบวนการเตรียมข้อมูลและการทำ Data Mining ที่เป็นต้นแบบของกระบวนการทาง Data Analysis และ Data Science ขึ้น ช่วงเวลานี้เองที่คำว่า “Data Science” ได้ถือกำเนิดและเริ่มเป็นที่รู้จักมากขึ้น

การขยายตัวของวงการข้อมูลยังไม่หยุดเพียงเท่านี้ ในปี ค.ศ. 2001 William S. Cleveland นักวิจัยทางสถิติจาก Bell Labs เขียนบทความ “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.” เสนอให้ขยายหลักสูตรในสาขาสถิติเพิ่มเป็นหลักสูตร “Data Science” เพื่อรองรับการใช้คอมพิวเตอร์ในการวิเคราะห์และสร้างตัวแบบและในช่วงเวลาเดียวกัน Leo Breiman นักสถิติชื่อดังจาก UC Berkeley ได้เขียนบทความคลาสสิกเกี่ยวกับความแตกต่างของสองวัฒนธรรมในการสร้างตัวแบบด้วยการใช้อัลกอริทึมกับตัวแบบทางสถิติที่ชื่อว่า Statistical Modeling: The Two Cultures แสดงให้เห็นว่าการนำข้อมูลมาใช้ในงานทำนายอนาคต (Prediction) กับการแปลความหมายของข้อมูล (Interpretation) ใช้ตัวแบบที่แตกต่างกันไป

กล่าวได้ว่าหลังจากช่วงปี ค.ศ. 2000 เป็นต้นมามีการพูดถึง Data Science บนสิ่งพิมพ์ต่างๆ มากขึ้นเรื่อย ๆ จนเมื่อปี ค.ศ. 2012 ที่ผ่านมา Harvard Business Review ได้ตีพิมพ์บทความ “Data Scientist: The Sexiest Job of the 21st Century” โดย Tom Davenport และ D.J. Patil ทำให้คำว่า Data Science ได้รับการพูดถึงจนเป็นกระแสนิยมถึงทุกวันนี้

Exploratory Data Analysis หรือ EDA เป็นกระบวนการตรวจสอบ สำรวจข้อมูลเบื้องต้น เป็นการวิเคราะห์ข้อมูลที่จำเป็นก่อนการนำข้อมูลไปใช้หรือนำไปวิเคราะห์เชิงลึก โดยประโยชน์ของการทำ EDA จะช่วยทำให้เราเข้าใจพื้นฐานของข้อมูลชุดนั้นๆ และเป็นการตรวจสอบความผิดพลาดของชุดข้อมูลได้อีกด้วย ถือว่าเป็นการวิเคราะห์ข้อมูลที่จำเป็นก่อนที่จะวิเคราะห์ข้อมูลแบบอื่น ๆ เช่นงาน Predictive ลักษณะการทำงานคือ สำรวจข้อมูลในมุมต่าง ๆ ในทุก ๆ ตัวแปรหรือเปรียบเทียบกันระหว่างตัวแปร EDA จะไม่มีการตั้งสมมุติฐานไว้ล่วงหน้าเหมือนการอนุมานทางสถิติ แต่จะให้ข้อมูลเป็นตัวบอกความเป็นไปของตัวมันเองแทน

แม้ว่าการทำ EDA จะเป็นการวิเคราะห์ขั้นต้นเพื่อทำความเข้าใจชุดข้อมูล แต่ผลลัพธ์ที่ได้จากการวิเคราะห์ สามารถทำให้เห็นถึงพฤติกรรมบางอย่างของข้อมูลนั้น ๆ ดังนั้นการทำ EDA จึงเป็นพื้นฐานที่จำเป็นและสำคัญอย่างมากในงานด้านการวิเคราะห์ข้อมูล (Data Analytics) โดยทั่วไปแล้ว Data Analyst และ Data Science ทั้งสองตำแหน่งสามารถทำ EDA ได้ อย่างไรก็ตามขั้นตอน ความละเอียด แนวทางและเครื่องมือที่ใช้ทำ EDA อาจจะแตกต่างกันได้

แม้ปัจจุบันจะมีเครื่องมือมากมายที่ทำให้การวิเคราะห์ข้อมูลขนาดใหญ่เกิดขึ้นได้ง่ายขึ้น แต่กระบวนการ EDA ก็ยังคงต้องอาศัยทักษะของนักวิเคราะห์เป็นส่วนใหญ่ ดังนั้นความรู้พื้นฐานทางสถิติจึงเป็นหัวใจสำคัญของการวิเคราะห์ ไม่ว่าจะขั้นพื้นฐานหรือขั้นสูงและนี้เองจึงทำให้นักวิเคราะห์ที่มีพื้นฐานทางด้านสถิติที่ดีจะสามารถปรับตัวให้เข้ากับเทคโนโลยีใหม่ ๆ ได้มากกว่านักวิเคราะห์ที่เน้นการใช้เครื่องมือแต่เพียงอย่างเดียว

บทที่ 1
ความรู้เบื้องต้นเกี่ยวกับการวิเคราะห์ข้อมูลเชิงสำรวจ

ในการวิเคราะห์สถิติเบื้องต้น โดยเฉพาะสถิติเชิงพรรณนา (Descriptive Statistics) ปกติแล้วเราจะใช้ค่าวัดแนวโน้มเข้าสู่ส่วนกลางเป็นค่าเฉลี่ยเลขคณิต (arithmetic mean) ซึ่งข้อมูลที่จะวิเคราะห์นั้น จะต้องเป็นข้อมูลเชิงปริมาณ แต่ค่าดังกล่าวมักจะมีความไว (sensitive) เมื่อเกิดค่าผิดปกติ ทำให้คุณสมบัติของค่าดังกล่าวด้อยลง แต่ก็สามารถแก้ไขปัญหานี้ได้ โดยการเปลี่ยนจากค่าเฉลี่ยเลขคณิตมาเป็นค่าฐานนิยม (mode) หรือค่ามัธยฐาน (median) และทำการวิเคราะห์ข้อมูลเหล่านั้นโดยใช้การวิเคราะห์ข้อมูลที่ไม่อิงพารามิเตอร์ (nonparametric) ซึ่งมักเรียกวิธีการดังกล่าวนี้ว่า การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis) โดยที่ค่าวัดแนวโน้มสู่ส่วนกลางที่ใช้ก็คือค่ามัธยฐาน (median) นั่นเอง

1. ความหมายของการวิเคราะห์ข้อมูลเชิงสำรวจ

การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis: EDA) เป็นเทคนิคที่ต้องการจะค้นหาเกี่ยวกับโครงสร้างของข้อมูล โดยวิธีที่ไม่เป็นทางการ เป็นวิธีแบบง่ายๆ เราทราบกันดีว่าสถิติอนุมาน ต้องการที่จะศึกษาข้อสมมติฐานเบื้องต้นเกี่ยวกับลักษณะทั่วไปของประชากรที่มีการแจกแจงแบบปรกติ และข้อมูลต้องเป็นอิสระกัน

2. กรอบของการวิเคราะห์ข้อมูลเชิงสำรวจ

2.1 วัตถุประสงค์

ต้องการที่จะค้นหาเกี่ยวกับโครงสร้างของข้อมูล

2.2 เป้าหมาย

การค้นหาและค้นพบโครงสร้างของข้อมูล วิธีการนี้ใช้เทคนิคการอนุมานเชิงสถิติ ซึ่งการอนุมานทางสถิติต้องการให้เราศึกษาสมมติฐานเกี่ยวกับธรรมชาติของประชากรและตัวอย่างที่มีการแจกแจงปรกติและมีความเป็นอิสระกัน

3. วิธีการวิเคราะห์ข้อมูลเชิงสำรวจ แบ่งออกเป็น 4 กลุ่ม ได้แก่

3.1 การทำให้เห็นเด่นชัด (Display)

เป็นการวิเคราะห์ด้วยแผนภาพลำต้นและใบ (stem-and-leaf) และแผนภาพกล่อง (Box plot) ซึ่งเป็นสื่อที่แสดงให้เราเห็นถึงโครงสร้างของแต่ละชุดข้อมูล

3.2 ความคลาดเคลื่อน หรือ ส่วนเศษเหลือ (Error or Residual)

เป็นการวิเคราะห์ความคลาดเคลื่อน หลังจากทำนายตัวแปรที่สนใจแล้ว จะทำการหาค่าความคลาดเคลื่อน ($e = \mathbf{Y}\text{-}\hat{\mathbf{Y}}$) ถ้าค่าคาดหมายของ e มีค่าเป็น 0 ก็แปลว่า ผลวิเคราะห์นั้นดีแล้ว

3.3 เรื่องการแสดงออก (Re-expression)

เป็นการแปลง (transform) ข้อมูลที่ใช้ประโยชน์ หากเราต้องการวิเคราะห์ข้อมูลโดยใช้สถิติเชิงอนุมาน (Inference Statistics) ซึ่งจำเป็นต้องตรวจสอบข้อสมมติ เช่นเมื่อข้อมูลไม่เป็นไปตามที่เราต้องการก็ต้องทำการแปลงข้อมูลก่อนการวิเคราะห์ ซึ่งเราสามารถตรวจสอบชุดข้อมูลที่มีอยู่ด้วยการวิเคราะห์ EDA ได้

3.4 ความต้านทาน (Resistance)

เป็นเรื่องคล้ายกับการวิเคราะห์การถดถอย (Regression) โดยที่ในเรื่องของการวิเคราะห์การถดถอยนั้น เราจะประมาณค่าพารามิเตอร์โดยใช้วิธีกำลังสองน้อยที่สุด (least square method) ซึ่งต้องตรวจสอบข้อสมมติให้ถูกต้องถึงจะมีความน่าเชื่อถือ แต่การหาเส้นต้านทาน (Resistance line) จะใช้มัธยฐาน (median) กับ ตำแหน่งความลึก (Depth) เป็นตัวอธิบาย มักวิเคราะห์โดยการทำซ้ำ (iterative) จนกว่าค่าความคลาดเคลื่อน (error) จะเปลี่ยนแปลงน้อยที่สุดหรือไม่เปลี่ยนแปลงเลย ปกติจะต้องใช้วิธีการเขียนโปรแกรมคำสั่งทางคอมพิวเตอร์ แต่ปัจจุบันนี้มีโปรแกรม Minitab สามารถใช้วิเคราะห์ EDA ได้ โดยใช้หลักการการปรับเรียบให้ตรง (Smooth) เป็นเครื่องมือช่วยในการวิเคราะห์

สำหรับการสรุปหรือการนำเสนอ การพรรณนาข้อมูลในแนวทางใหม่ คือ แทนที่จะอธิบายเฉพาะค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐาน ในปัจจุบันมีการนำเสนอค่าที่สำคัญ 5 ค่า ที่เรียกว่าตัวเลขสรุป 5 ค่า (Five Numbers Summary) ซึ่งประกอบด้วย ค่าน้อยที่สุด (Smallest value) ค่าควอร์-ไทล์ที่ 1 (1^{st} Quartile) หรือเปอร์เซ็น-ไทล์ที่ 25 (25^{th} Percentile) ค่าเฉลี่ยเลขคณิต (Arithmetic Median) ค่าควอร์-ไทล์ ที่ 3 (3^{rd} Quartile) หรือเปอร์เซ็น-ไทล์ ที่ 75 (75^{th} Percentile) และ ค่ามากที่สุด (Largest value)