บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัญหาประการหนึ่งในการวิเคราะห์ข้อมูลในทางสถิติคือ การที่ข้อมูลหรือค่าสังเกตที่เก็บรวบรวมได้ไม่เป็นไปตามสภาวการณ์ที่กำหนดหรือควบคุมไว้ เช่น การทดลองทางการแพทย์ ด้านชีววิทยา เป็นต้น ทำให้ข้อมูลหรือค่าสังเกตบางค่าสูงหรือต่ำมาก หรือเป็นค่าสังเกตที่ไม่ได้มาจากประชากรเดียวกับค่าสังเกตส่วนใหญ่ โดยกรณีหลังนี้จะพบมากในข้อมูลที่ได้จากการวางแผนการทดลองผิดพลาด ข้อมูลที่มีลักษณะดังกล่าวจะเรียกว่าเป็น "**ค่าผิดปรกติ**" (outliers)

ในปี ค.ศ. 1960 อังโคมบ์ (Ancombe) กล่าวว่าสาเหตุของข้อมูลผิดปรกติมี 3 ประการ ได้แก่

1. ความผันแปรที่มีอยู่ในประชากรที่ศึกษา (inherent variability) เป็นความ ผันแปรที่ไม่สามารถหลีกเลี่ยงได้ แม้จะมีการควบคุมหรือการวัดหรือการปฏิบัติเป็นอย่างดีก็ยังมีอยู่ และแก้ไขไม่ได้นอกจากจะเปลี่ยนประชากรหรือวัตถุประสงค์ในการศึกษา

2. ความคลาดเคลื่อนที่เกิดจากการวัด (measurement error) ความผิดพลาดชนิดนี้เกิดจากการบันทึกข้อมูลหรือเครื่องมือเครื่องใช้ในการวัดมีคุณภาพต่ำ ความคลาดเคลื่อนนี้อาจแก้ไขให้หมดไปได้

3. ความคลาดเคลื่อนที่เกิดจากการปฏิบัติการ (execution error) เช่น การลงรหัส การเจาะบัตร เป็นต้น ความผิดพลาดชนิดนี้สามารถลดลงได้ถ้ามีการระมัดระวัง ป้องกันไว้ก่อน

ในปี ค.ศ. 1980 ฮอกินส์ (Hawkins) ให้คำจำกัดความของค่าผิดปรกติว่า "เป็นค่าสังเกตที่เบี่ยงเบนไปจากค่าสังเกตอื่นอย่างมากจนทำให้สงสัยว่าค่าสังเกตนั้นได้มาจากวิธีการ

(mechanism) อื่น" และในปี 1983 เบกแมนและคุก (Beckman and Cook) ได้สรุปความหมายของค่าผิดปรกติออกเป็น 2 ลักษณะคือ

ก) ค่าสังเกตที่มีค่าสูงหรือต่ำมาก (extreme) หรือเป็นค่าที่เบี่ยงเบนไปจากกลุ่มที่ศึกษา เรียกค่าผิดปรกตินี้ว่า **discordant observation**

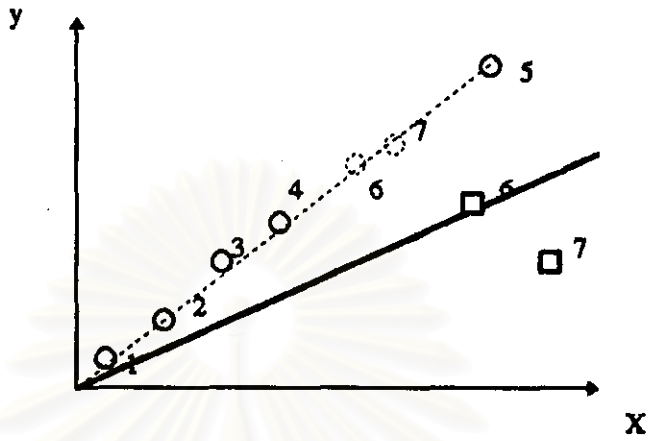
ข) ค่าสังเกตที่มีลักษณะการแจกแจงแตกต่างจากลักษณะการแจกแจงของประชากรที่สนใจศึกษาและเรียกค่าผิดปรกตินี้ว่า **contaminate observation**

การวิเคราะห์ความถดถอยเป็นวิธีการวิเคราะห์ข้อมูลทางสถิติที่มีการนำไปใช้ในงานวิจัยทางด้านสังคมศาสตร์และวิทยาศาสตร์ โดยการนำข้อมูลที่สังเกตหรือเก็บรวบรวมได้มาใช้ประมาณหรือพยากรณ์สิ่งที่สนใจ ข้อมูลที่ใช้ในการวิเคราะห์ความถดถอยบางครั้งอาจมีค่าผิดปรกติปนอยู่ ทำให้ผลการวิเคราะห์ความถดถอยเกิดความผิดพลาดขึ้นได้ ทั้งนี้เพราะค่าผิดปรกติดังกล่าวมีผลต่อการประมาณค่าสัมประสิทธิ์การถดถอย ($\underset{\sim}{\beta}$) ซึ่งทำให้สมการถดถอยที่ได้เบี่ยงเบนไปในทิศทางหรือตำแหน่งของค่าผิดปรกติ การเบี่ยงเบนไปในทิศทางหรือตำแหน่งของค่าผิดปรกติดังกล่าวทำให้เกิดปัญหาในการตรวจสอบค่าผิดปรกติเกิดขึ้น 2 ชนิด คือ **การตรวจสอบค่าผิดปรกตินั้นไม่พบ และการตรวจสอบพบค่าผิดปรกติแต่ค่าผิดปรกติที่ตรวจพบไม่ใช่ค่าผิดปรกติจริงในข้อมูล** กล่าวคือ หลังจากที่ทำการวิเคราะห์ความถดถอยด้วยวิธีกำลังสองน้อยสุดแล้วค่าตัวแปรตามที่พยากรณ์ได้จะมีความคลาดเคลื่อนมากน้อยแตกต่างกัน **ค่าผิดปรกติในการวิเคราะห์ความถดถอยหมายถึง ค่าสังเกตที่มีความคลาดเคลื่อนสูง** แต่เนื่องจากค่าผิดปรกติที่มีในข้อมูลดึงสมการการถดถอยเบี่ยงเบนไปในทิศทางหรือตำแหน่งของมัน ก็จะส่งผลให้ความคลาดเคลื่อนของมันต่ำ ทำให้ตรวจสอบไม่พบค่าผิดปรกติดังกล่าวหรืออาจส่งผลให้ค่าสังเกตที่ดีหรือสะอาด* (good or clean observation) มีความคลาดเคลื่อนสูงจนกลายเป็นค่าผิดปรกติได้เมื่อตรวจสอบด้วยวิธีพื้นฐาน

---

* ข้อมูลดีหรือสะอาด หมายถึง ข้อมูลนั้นเป็นข้อมูลปรกติ

ปัญหาที่เกิดขึ้น เรียกว่า **มาซคคิงเอฟเฟค** * (masking effect) และ **ชวอมพิงเอฟเฟค** ** (swamping effect) ตามลำดับ ลักษณะการเกิดมาซคคิงและชวอมพิงเอฟเฟคแสดงในรูปที่ 1.1



รูปที่ 1.1 : แสดงลักษณะของการเกิดมาซคคิงเอฟเฟค ณ ตำแหน่งที่ 6 และ ชวอมพิงเอฟเฟค ณ ตำแหน่งที่ 5

จากรูปที่ 1.1 ข้อมูลที่เรานำมาวิเคราะห์มีความสัมพันธ์เชิงเส้นตรงด้วยเส้นประ แต่ถ้าค่าที่ 6 และ 7 เบี่ยงเบนจากข้อมูลโดยส่วนใหญ่เราจะได้ความสัมพันธ์เชิงเส้นตรงด้วยเส้นทึบ *** เมื่อเราทำการตรวจสอบค่าผิดปรกติด้วยวิธีพื้นฐาน คือ พิจารณาค่าที่มีความคลาดเคลื่อนสูงเป็นค่าผิดปรกติและใช้ทุกค่าสังเกตในการวิเคราะห์ เราจะพบว่าค่าที่ 6 จะมีความคลาดเคลื่อนต่ำทำให้ตรวจสอบค่าผิดปรกติดังกล่าวไม่พบ แต่ค่าที่ 5 จะมีความคลาดเคลื่อนสูงทำให้ค่าที่ 5 กลายเป็นค่าผิดปรกติที่ถูกตรวจพบ ดังนั้นการตรวจสอบค่าผิดปรกติจำเป็นต้องหาวิธีที่มีความแกร่ง (robust) พอที่จะไม่ทำให้เกิดปัญหาทั้ง 2 ชนิด หรือยอมให้เกิดได้น้อยที่สุด เพื่อจะได้ทำให้พบค่าผิดปรกติจริงๆ และจะได้แก้ปัญหาของข้อมูลมีค่าผิดปรกติก่อนที่จะนำไปใช้ในการวิเคราะห์ความถดถอยเพื่อทำให้ได้ผลการพยากรณ์ทางสถิติซึ่งถูกต้องต่อไป

---

* เหตุการณ์ที่ค่าผิดปรกติค่าหนึ่งมีผลต่อค่าผิดปรกติอีกค่าหนึ่งทำให้ไม่สามารถตรวจพบค่าผิดปรกตินั้น

** เหตุการณ์ที่ค่าผิดปรกติมีผลต่อค่าสังเกตอื่นๆ ที่ไม่ใช่ค่าผิดปรกติ ทำให้ค่านั้นกลายเป็นค่าผิดปรกติซึ่งถูกตรวจพบ

*** การเบี่ยงเบนของค่าที่ 6 และ 7 อาจทำให้เรามองความสัมพันธ์ระหว่าง x กับ y ในรูปโพลิโนเมียลได้ แต่การใช้ความถดถอยโพลิโนเมียลให้ผลการประมาณค่าสัมประสิทธิ์ความถดถอยไม่แม่นยำเท่าการถดถอยเชิงเส้นตรงเนื่องจากแต่ละพจน์ของตัวแปรอิสระมีความสัมพันธ์กันเอง และข้อมูลที่เรากำลังศึกษาก็มีความสัมพันธ์เชิงเส้นระหว่าง x กับ y จึงไม่เหมาะสมที่จะใช้การถดถอยโพลิโนเมียล

วิธีการตรวจสอบค่าผิดปรกติในการวิเคราะห์ความถดถอยเชิงเส้นได้มีผู้ศึกษาไว้ดังนี้ ในปี ค.ศ. 1967 มิกคีย์ (Mickey) ดัน (Dun) และ คลาซค (Clask) ได้ใช้วิธีการวิเคราะห์ความถดถอยแบบขั้นตอน (Stepwise Regression) และการเพิ่มตัวแปรหุ่น (Dummy Variable) เข้าไปในสมการการถดถอยเพื่อทำการแยกค่าผิดปรกติ แต่วิธีนี้ไม่เหมาะสมในกรณีที่มีค่าผิดปรกติมากกว่า 1 ค่า ในปี ค.ศ. 1985 เมอร์วิน จี มาราซิง (Merveyn G. Marasinghe) ได้เสนอวิธีการวิเคราะห์แบบหลายขั้นตอน (multistage procedure) และ ตัวสถิติ F_k^* เพื่อใช้ตรวจสอบค่าผิดปรกติหลายค่าในการวิเคราะห์ความถดถอยเชิงเส้น ในปี ค.ศ. 1990 รูซไซ(Rousseeue) และ แวน ไซเมอร์เรน(van Zomeren) ได้นำตัวประมาณที่ได้จากทรงรีที่มีปริมาตรต่ำสุด (minimum volume elipsoid (MVE)) มาหาระยะทางที่แกร่ง (robust distance) เพื่อใช้ในการตรวจสอบค่าผิดปรกติแบบหลายตัวแปร (multvariate outliers) แต่การหาตัวประมาณ MVE นั้นจะต้องหาเซตย่อยที่มีจำนวนสมาชิก(ค่าสังเกต) อย่างน้อยครึ่งหนึ่งของข้อมูล ถ้าให้จำนวนสมาชิกอย่างน้อยครึ่งหนึ่งของข้อมูลเท่ากับ h และ ขนาดตัวอย่างเท่ากับ n จะต้องทำการคำนวณหาปริมาตรของทรงรีเท่ากับ $\binom{n}{h}$ ปริมาตรแล้วจึงจะเลือกทรงรีที่มีปริมาตรต่ำที่สุด ดังนั้นจึงทำให้เสียค่าใช้จ่ายและเวลาในการคำนวณสูงและบางครั้งการคำนวณอาจเป็นไปไม่ได้ ต่อมา ในปี ค.ศ. 1992 ฮาได (Hadi) เสนอวิธีการหา MVE ซึ่งจะทำให้ได้ MVE เพียงค่าเดียว (uniqe) และยังใช้ได้ดีแม้ว่าเมทริกซ์ความแปรปรวนร่วมจะเป็นเมทริกซ์เอกฐาน (singular metrix) ก็ตาม ฮาไดได้แก้ปัญหาที่เมทริกซ์ความแปรปรวนร่วมเป็นเมทริกซ์เอกฐาน โดยการถ่วงน้ำหนักเวกเตอร์เจาะจง (eigenvecters) ด้วยค่าเจาะจงสูงสุด (maximum eigenvalues) ของเมทริกซ์ความแปรปรวนร่วม และในปี ค.ศ. 1994 ฮาได (Hadi) ก็ได้แก้ปัญหาดังกล่าวอีกครั้ง โดยการรวมค่าสังเกตที่อยู่ถัด ๆ ไปเข้าไปในเซตย่อยเท่าที่จำเป็น จนกว่าเมทริกซ์ความแปรปรวนร่วมจะเป็นเมทริกซ์ไม่เอกฐาน (non-singular) ในปี ค.ศ. 1993 ฮาได(Hadi) และไซมันนอฟฟ์(Simonoff) ได้เสนอวิธีการตรวจสอบค่าผิดปรกติหลายค่า (multiple outliers) ในตัวแบบเชิงเส้นโดยอาศัยแนวคิดของฮาได ในปี ค.ศ. 1992 ไดนิฟาร์ด (Kianifard) และซวอลโล(Swallow) ได้เสนอวิธีการตรวจสอบค่าผิดปรกติโดยใช้เกณฑ์ความคลาดเคลื่อนเวียนเกิด (recursive residual) และได้เสนอตัวสถิติทดสอบ 2 วิธี ได้แก่ วิธีเวียนเกิดโดยลำดับ (sequential recursive method) และวิธีเวียนเกิดดัดแปร (modified recursive method) ในปี พ.ศ. 2532 บุญสม หรรษศิริพจน์ ได้ศึกษาวิธีตรวจสอบค่าผิดปรกติในสมการการถดถอยเชิงพหุ โดยศึกษาเปรียบเทียบวิธีของ เดนนิส คุก (Dennis Cook, 1977), วิธีของแอนดรูและเพรตจิบอน (Andrew and Pregibon, 1978) และวิธีของ จี แบร์รี (G. Barre Weterill, 1986) ซึ่งศึกษาในกรณีที่การแจกแจงของความคลาดเคลื่อนมี 2 แบบคือ การแจกแจงปรกติปลอมปนในตำแหน่งและการแจกแจงปรกติปลอมปนในสเกลผลการศึกษาปรากฏว่าวิธีของ จี แบร์รี และ วิธีของแอนดรูและเพรตจิบอน มีความสามารถในการควบคุมความผิดพลาดประเภทที่ 1 ได้ดี ในปี

พ.ศ. 2533 สมชาย รัตนเลิศนุสรณ์ ได้ศึกษาวิธีการตรวจสอบค่าผิดปรกติในการวิเคราะห์ความถดถอยเชิงเส้น โดยการเปรียบเทียบวิธีของ จี แบร์รี (G. Barre Weterill, 1986) วิธีของทิดเจน, มัวร์และเบกแมน (Tietjen, Moore and Beckman, 1973) และวิธีของ เมอร์วิน จี มาราซิง (Mervyn G.Marasinghe,1985) โดยศึกษาในกรณีที่การแจกแจงของความคลาดเคลื่อนมี 2 ลักษณะ คือ การแจกแจงแบบหางยาวกว่าปรกติ ได้แก่ การแจกแจงปรกติปลอมปนในสเกล การแจกแจงปรกติปลอมปนในตำแหน่ง และการแจกแจงที ส่วนการแจกแจงแบบแบ้ขวา ได้แก่ การแจกแจงลอกนอร์มอล แกมมา และไวบูลล์ พบว่าวิธีของ จี แบร์รี และวิธีของเมอร์วิน จี มาราซิง ควบคุมความคลาดผิดพลาดประเภทที่ 1 ได้ดีใกล้เคียงกัน ส่วนวิธีของทิดเจน, มัวร์และเบกแมนควบคุมได้น้อย

เนื่องจากตัวสถิติทดสอบแต่ละตัวมีความสามารถในการตรวจสอบค่าผิดปรกติได้ต่างกัน บางวิธีอาจทำให้เกิดมาซคคิงและซวอมพิงเอฟเฟ็คได้เป็นผลให้ตรวจค่าผิดปรกติจริงไม่พบ หรือพบไม่ครบทุกค่า ผู้วิจัยจึงสนใจที่จะเปรียบเทียบวิธีการตรวจสอบค่าผิดปรกติในการวิเคราะห์ความถดถอย เชิงเส้นโดยเปรียบเทียบวิธีของ เมอร์วิน จี มาราซิง (Mervyn G. Marasinghe, 1985) วิธีของไคนิฟาร์ดและซวอลโล (Kianifard and Swallow, 1990) และวิธีของฮาไดและไซมันนอฟฟ์ (Hadi and Simonoff,1993)

1.2 วัตถุประสงค์

เพื่อศึกษาเปรียบเทียบตัวสถิติที่ใช้ตรวจสอบข้อมูลผิดปรกติ ในการวิเคราะห์ความถดถอยเชิงเส้น ของตัวสถิติทดสอบ 4 ตัว คือ

1. ตัวสถิติทดสอบของเมอร์วิน จี มาราซิง (MV)
2. ตัวสถิติทดสอบของฮาไดและไซมันนอฟฟ์ (HS)
3. ตัวสถิติทดสอบของไคนิฟาร์ดและสวอลโล ซึ่งมี 2 วิธี คือ
   3.1 วิธีเวียนเกิดโดยลำดับ (sequential recursive method(SRM))
   3.2 วิธีเวียนเกิดดัดแปร (modified recursive method(MRM))

ในการเปรียบเทียบตัวสถิติทั้ง 4 ตัว ผู้วิจัยจะพิจารณาจากเกณฑ์ความสามารถในการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1, ความน่าจะเป็นซึ่งค่าผิดปรกติที่ถูกตรวจพบเป็นค่าผิดปรกติจริงทุกค่า (p_1), ความน่าจะเป็นซึ่งทำให้เกิดมาซคคิงเอฟเฟ็ค (p_2) และ ความน่าจะเป็นซึ่งทำให้เกิดซวอมพิงเอฟเฟ็ค (p_3)

1.3 สมมติฐาน

ตัวสถิติทดสอบ HS มีประสิทธิภาพดีที่สุด (มีค่าของ p_1 สูงกว่าตัวอื่น หรือมีค่า p_2 ต่ำกว่าตัวอื่น หรือมีค่า p_3 ต่ำกว่าตัวอื่น) ในการตรวจพบค่าผิดปรกติหลายค่า เพราะว่าการหาค่าผิดปรกติของวิธีการนี้ มีการคัดเลือกข้อมูลที่ผิดปรกติออกจากค่าดีอย่างเคร่งครัดก่อนจะทำการทดสอบ และการทดสอบค่าผิดปรกติจะทดสอบจากค่าปลายน้อยสุดไปหาค่าปลายมากสุด (least to most extreme value) ซึ่งโอกาสที่เราจะพบค่าผิดปรกติจริงมีมาก

1.4 ข้อตกลงเบื้องต้น

1. ความคลาดเคลื่อน (error) มีการแจกแจงแบบเดียวกัน (ยกเว้นกรณีค่าผิดปรกติค่าความคลาดเคลื่อนอาจจะมีการปลอมปนของการแจกแจงอื่นได้และอิสระกัน)

2. การสร้างค่าผิดปรกติเราจะกำหนดตำแหน่งของค่าผิดปรกติเพื่อหาค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ค่า p_1 , p_2 และ p_3

3. การประมาณค่าสัมประสิทธิ์ความถดถอย ($\hat{\beta}$) จะใช้วิธีกำลังสองน้อยสุด (Least Square method)

1.5 ขอบเขตการวิจัย

1. ศึกษาวิธีการตรวจสอบค่าผิดปรกติในการวิเคราะห์ความถดถอยเชิงเส้น โดยกำหนดจำนวนตัวแปรอิสระเท่ากับ 1,3 และ 5 ตัว

2. ศึกษากรณีขนาดตัวอย่างเท่ากับ 20, 50 และ 100

3. ศึกษาโดยกำหนดจำนวนค่าผิดปรกติเท่ากับ 0, 1, 2 และ 3 ค่า

4. กำหนดระดับนัยสำคัญ (α) เท่ากับ 0.01 และ 0.05

5. การแจกแจงของความคลาดเคลื่อนที่ศึกษามีดังนี้

5.1 <u>การแจกแจงหางยาวกว่าการแจกแจงปรกติ</u> ได้แก่

5.1.1 <u>การแจกแจงปรกติปลอมปนในสเกล</u>* (Scale - contaminated normal distribution) เราจะศึกษาในกรณีที่ค่าพารามิเตอร์แสดงสเกล (c) เท่ากับ 3, 5 และ 10 ตามลำดับ และมีร้อยละการปลอมปนเท่ากับร้อยละของค่าผิดปรกติที่กำหนดในแต่ละขนาดตัวอย่าง ดังตารางที่ 1.1

5.1.2 <u>การแจกแจงแบบปรกติปลอมปนในตำแหน่ง</u>** (location contaminated normal distribution) เราจะศึกษาในกรณีที่ค่าพารามิเตอร์แสดงตำแหน่ง (a) เท่ากับ 3, 5 และ 10 ตามลำดับ และมีร้อยละการปลอมปนเท่ากับร้อยละของค่าผิดปรกติที่กำหนดในแต่ละขนาดตัวอย่าง ดังตารางที่ 1.1

ตารางที่ 1.1 แสดงร้อยละการปลอมปนของการแจกแจงปรกติปลอมปนในตำแหน่ง และการแจกแจงปรกติปลอมปนในสเกล

จำนวนค่าผิดปรกติ	ขนาดตัวอย่าง		
	20	50	100
1	5%	2%	1%
2	10%	4%	2%
3	15%	6%	3%

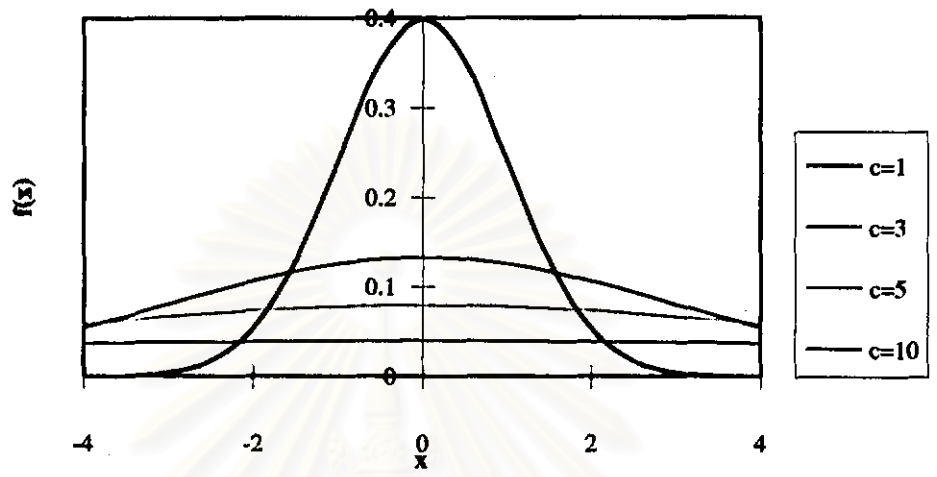
5.1.3 <u>การแจกแจงที</u> (t-distribution) เราจะศึกษากรณีที่ขนาดตัวอย่างเท่ากับ 20 ณ ระดับขั้นความเสรี (d.f.) เท่ากับ 14, 16 และ 18 ของจำนวนตัวแปรอิสระเท่ากับ 1, 3 และ 5 ตามลำดับ เนื่องจากเมื่อขนาดตัวอย่างมากๆ การแจกแจงทีจะลู่เข้าสู่การแจกแจงปรกติ และทำให้ความคลาดเคลื่อนต่ำกว่าความเป็นจริง

---

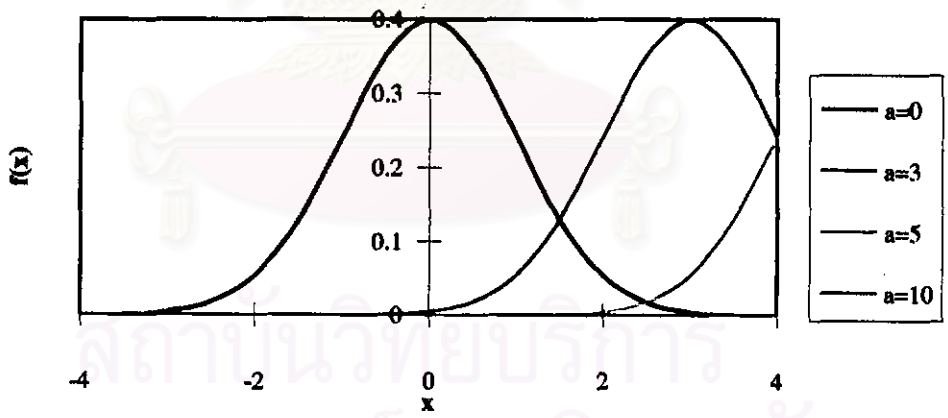
* ผู้วิจัยทดลองศึกษาเมื่อ $\mu = 0$ ด้วยค่าพารามิเตอร์แสดงสเกล(c)ต่างๆ พบว่าลักษณะโค้งการแจกแจงเปลี่ยนแปลงไปตามค่า c ดังรูปที่ 1.2 ซึ่งแสดงให้เห็นว่าเมื่อค่า c มากๆ โอกาสที่จะเกิดค่าผิดปรกติมีมากขึ้น

** ผู้วิจัยทดลองศึกษาเมื่อ $\sigma^2 = 1$ ด้วยค่าพารามิเตอร์แสดงตำแหน่ง(a)ต่างๆ พบว่าลักษณะโค้งการแจกแจงเปลี่ยนแปลงไปตามค่า a ดังรูปที่ 1.3 ซึ่งแสดงให้เห็นว่าเมื่อค่า a มากๆ โอกาสที่จะเกิดค่าผิดปรกติมีมากขึ้น

รูปที่ 1.2 แสดงการแจกแจงปรกติปลอมปนในสเกล



รูปที่ 1.3 แสดงการแจกแจงปรกติปลอมปนในตำแหน่ง



5.2 <u>การแจกแจงแบบเบ้</u> * (skewed distribution) เราจะศึกษาการแจกแจงดังนี้

5.2.1 <u>การแจกแจงลอกนอร์มอล</u> ** (lognormal distribution) ผู้วิจัยจะศึกษาในกรณีที่ค่า $\mu = 0$, $\sigma^2 = 0.1, 0.2$ และ 0.7 การแจกแจงลอกนอร์มอลมีฟังก์ชันความหนาแน่นดังนี้

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(\log(x) - \mu)^2}{\sigma^2} \right] \; ; \; -x > 0, \sigma > 0, -\infty < \mu < \infty$$

โดยที่ μ และ σ^2 คือค่าเฉลี่ยและความแปรปรวนของ y เมื่อ $y = \log(x)$ มีการแจกแจงปรกติ

ค่าเฉลี่ยและความแปรปรวนของ x มีดังนี้

$$E(x) = \exp\left(\mu + \frac{1}{2}\sigma^2 \right)$$

$$Var(x) = \exp(2\mu + \sigma^2) * [\exp(\sigma^2) - 1]$$

5.2.2 <u>การแจกแจงแกมมา</u> *** (gamma distribution) ผู้วิจัยจะศึกษากรณีที่ค่าพารามิเตอร์แสดงตำแหน่ง $\beta = 1$ และค่าพารามิเตอร์แสดงรูปร่าง $\alpha = 1, 2, 3$ และ 10 การแจกแจงแกมมามีฟังก์ชันความหนาแน่นดังนี้

$$f(x) = \frac{x^{\alpha-1} \exp(-x/\beta)}{\beta^{\alpha}\Gamma(\alpha)} \; ; x > 0, \alpha > 0, \beta > 0$$

โดยที่ β เป็นพารามิเตอร์แสดงตำแหน่ง

และ α เป็นพารามิเตอร์แสดงรูปร่าง

ค่าเฉลี่ย และความแปรปรวนของตัวแปร x คือ

$$E(x) = \alpha\beta$$

$$Var(x) = \alpha\beta^2$$

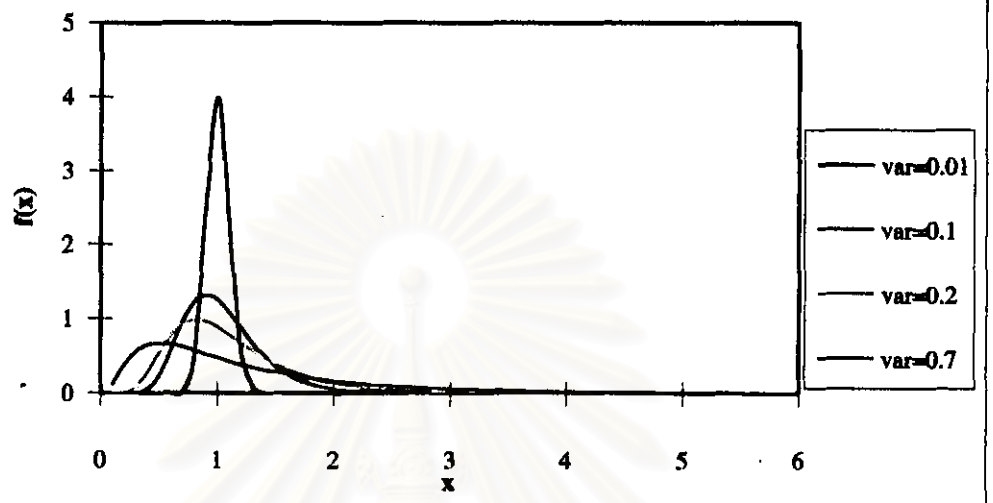
---

* ศึกษาเฉพาะการแจกแจงเบ้ขวา เนื่องจากการแจกแจงเบ้ซ้ายเป็นส่วนกลับของการแจกแจงเบ้ขวา

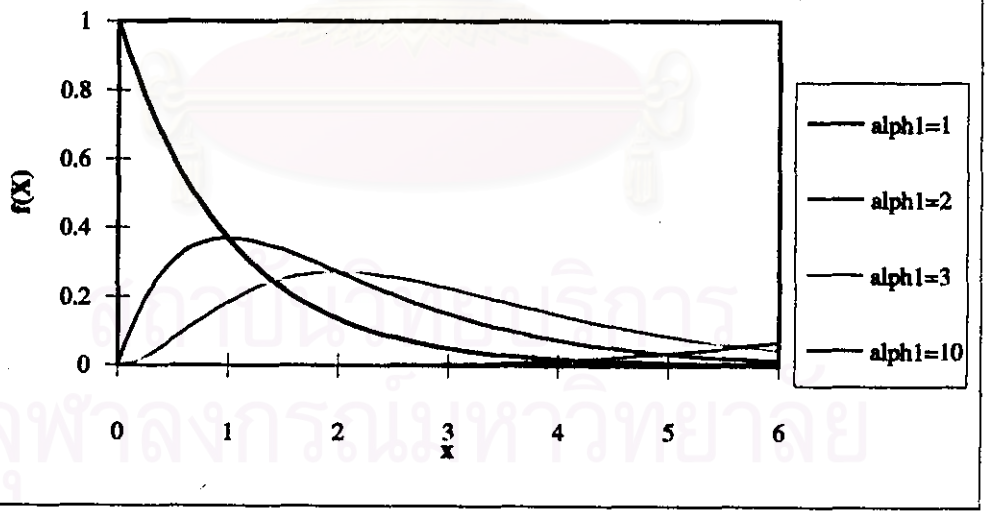
** เลือกค่า σ^2 ตามลักษณะโค้งการแจกแจงที่เปลี่ยนไปดังรูปที่ 1.4

*** เลือกค่า α ตามลักษณะโค้งการแจงแจงที่เปลี่ยนไปดังรูปที่ 1.5

รูปที่ 1.4 แสดงโค้งการแจกแจงฉอกนอร์มอล



รูปที่ 1.5 แสดงโค้งการแจกแจงแกมมา



5.2.3 การแจกแจงไวบูลล์* (Weibull distribution) ผู้วิจัยจะศึกษากรณีที่ค่าพารามิเตอร์แสดงตำแหน่ง $\beta = 1$ และค่าพารามิเตอร์แสดงรูปร่าง $\alpha = 1, 2, 3$ และ 10 การแจกแจงไวบูลล์มีฟังก์ชันความหนาแน่นดังนี้

$$f(x) = \frac{\alpha x^{\alpha-1} \exp\left[-(x/\beta)^{\alpha}\right]}{\beta^{\alpha}} \quad ; x>, \alpha > 0, \beta > 0$$

โดยที่ β เป็นพารามิเตอร์แสดงตำแหน่ง

และ α เป็นพารามิเตอร์แสดงรูปร่าง

ค่าเฉลี่ย และความแปรปรวนของตัวแปร x คือ

$$E(x) = \beta\, \Gamma\left(1+\frac{1}{\alpha}\right)$$

$$Var(x) = \beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$$

6. ทำการสร้างแบบจำลองข้อมูลให้มีสถานการณ์ตามต้องการ โดยใช้วิธีมอนติคาร์โลด้วยเครื่อง IBM/3031 ณ สถาบันบริการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย และเขียนโปรแกรมด้วยภาษา FORTRAN

1.6 คำจำกัดความ

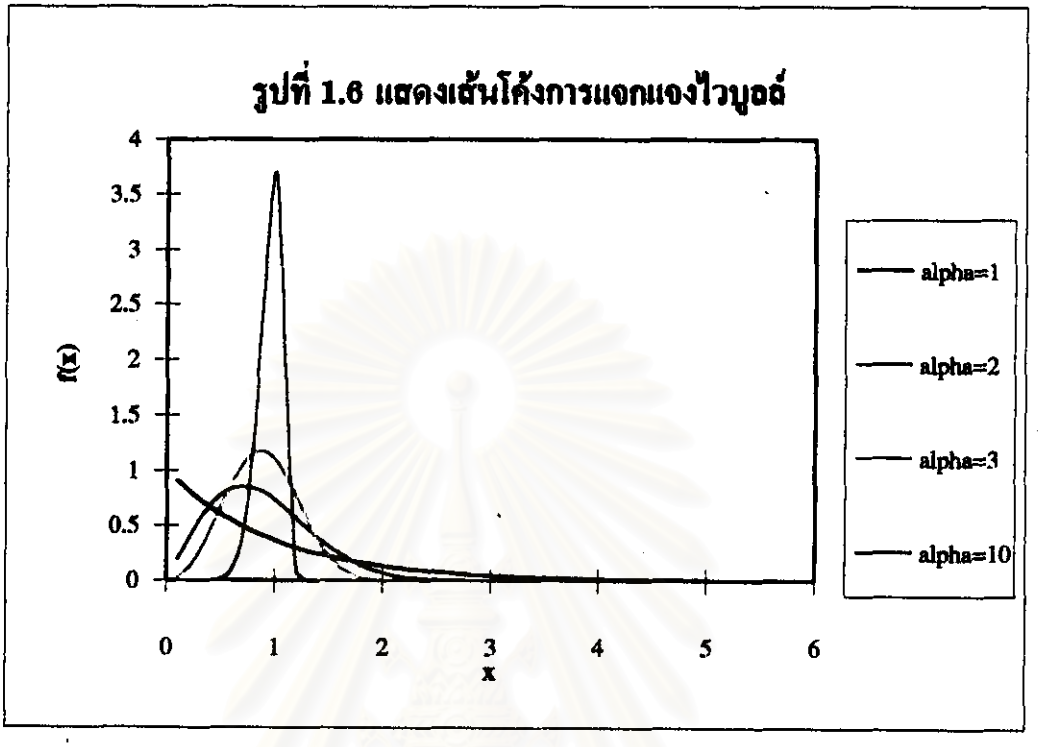
1. ค่าผิดปรกติ (Outliers) หมายถึงค่าสังเกตที่มีค่ามากกว่าหรือน้อยกว่าค่าสังเกตอื่นๆ หรือค่าสังเกตที่ไม่ได้มาจากประชากรเดียวกัน

2. มาซคกิงเอฟเฟ็ค (Masking effect) หมายถึง เหตุการณ์ที่ค่าผิดปรกติค่าหนึ่งมีผลต่อค่าผิดปรกติอีกค่าหนึ่งทำให้ไม่สามารถตรวจพบค่าผิดปรกตินั้น

3. ซวอมพิงเอฟเฟ็ค (Swamping effect) หมายถึง เหตุการณ์ที่ค่าผิดปรกติมีผลต่อค่าสังเกตอื่นๆ ที่ไม่ใช่ค่าผิดปรกติ ทำให้ค่านั้นกลายเป็นค่าผิดปรกติซึ่งถูกตรวจพบ

---

* เลือกค่า α ตามลักษณะโค้งการแจงแจงที่เปลี่ยนไปดังรูปที่ 1.6



รูปที่ 1.6 แสดงเส้นโค้งการแจกแจงไวบูลล์

4. ความผิดพลาดประเภทที่ 1 (type I error) หมายถึง ความคลาดเคลื่อนที่เกิดจากการปฏิเสธสมมติฐานว่าง (null hypothesis) เมื่อสมมติฐานว่างเป็นจริง

5. ความผิดพลาดประเภทที่ 2 (type II error) หมายถึง ความคลาดเคลื่อนที่เกิดจากการยอมรับสมมติฐานว่างเมื่อสมมติฐานว่างไม่จริง

6. อำนาจการทดสอบ (power of the test) หมายถึง ความน่าจะเป็นที่จะปฏิเสธสมมติฐานว่างเมื่อสมมติฐานว่างไม่จริง

7. ความน่าจะเป็นซึ่งค่าผิดปรกติที่ถูกตรวจพบเป็นค่าผิดปรกติจริงทุกค่า* (p_1) หมายถึง ความน่าจะเป็นซึ่งตัวสถิติทดสอบจะตรวจพบค่าผิดปรกติโดยที่ค่าผิดปรกตินั้นเป็นค่าผิดปรกติจริงที่เรากำหนดทุกค่า

8. ความน่าจะเป็นซึ่งทำให้เกิดมาซกกิงเอฟเฟ็ค* (p_2) หมายถึง ความน่าจะเป็นที่จะไม่พบค่าผิดปรกติทั้งที่ในข้อมูลที่ทำการทดสอบนั้นมีค่าผิดปรกติอยู่

9. ความน่าจะเป็นซึ่งทำให้เกิดชวอมพิงเอฟเฟ็ค* (p_3) หมายถึงความน่าจะเป็นที่จะตรวจพบค่าผิดปรกติทั้งที่ในข้อมูลนั้นไม่มีค่าผิดปรกติ หรือความน่าจะเป็นที่จะตรวจพบค่าผิดปรกติโดยที่ค่าผิดปรกตินั้นเป็นค่าผิดปรกติไม่จริง (เป็นค่าปรกติ)

10. ความแกร่งของการทดสอบ (Robustness) หมายถึง คุณสมบัติของการทดสอบที่ไม่ไวต่อการเปลี่ยนแปลงของปัจจัยอื่น ที่ไม่ใช่ปัจจัยที่ต้องการทดสอบ เช่น การฝ่าฝืนข้อตกลงเบื้องต้นของการทดสอบนั้น ซึ่งสิ่งที่ใช้พิจารณาความแกร่งของการทดสอบ คือ ความผิดพลาดประเภทที่ 1

1.7 เกณฑ์การตัดสินใจ

เกณฑ์การตัดสินใจว่าวิธีการตรวจสอบค่าผิดปรกติวิธีใดมีประสิทธิภาพสูงสุดจะพิจารณาภายใต้สมมติฐานว่าง

H : ไม่มีค่าผิดปรกติ

เทียบกับ K : มีค่าผิดปรกติอย่างน้อย 1 ค่า

---

* เกณฑ์ที่ใช้วัดดังกล่าวในข้อ 8, 9 และ 10 เดวิดและพอลสัน (David and Paulson) ได้เสนอไว้ในปีค.ศ. 1965 ต่อมาในปี ค.ศ.1990 ไคนิฟาร์ดและชวอลโล (Kianifard and Swollow) และ ในปี ค.ศ. 1993 ฮาได (Hadi) ได้นำมาดัดแปลงใช้ในการวัดประสิทธิภาพของวิธีการรวจสอบค่าผิดปรกติ

ผู้วิจัยจะทำการเปรียบเทียบความน่าจะเป็นของตัวสถิติทดสอบ 2 ลักษณะ ดังนี้

1. ความสามารถในการควบคุมความน่าจะเป็นความผิดพลาดประเภทที่ 1 ของการทดสอบแต่ละสถานการณ์ โดยใช้เกณฑ์ของคอคแรน (Cochran) และบราดเลย์ (Bradley)* ซึ่งเราจะพิจารณาดังนี้

ก) <u>เกณฑ์ของคอคแรน</u> (Cochran) เราจะพิจารณาว่าตัวสถิติทดสอบใดมีค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ซึ่งได้จากการทดลอง อยู่ในช่วง [0.007, 0.015] และ [0.04, 0.06] ณ ระดับนัยสำคัญ 0.01 และ 0.05 ตามลำดับ เราจะถือว่าตัวสถิติทดสอบนั้นควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ได้

ข) <u>เกณฑ์ของบราดเลย์</u> (Bradley) เราจะพิจารณาว่าตัวสถิติทดสอบใดมีค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ซึ่งได้จากการคำนวณ อยู่ในช่วง [0.005, 0.015] และ [0.025, 0.075] ณ ระดับนัยสำคัญ 0.01 และ 0.05 ตามลำดับ เราจะถือว่าวิธีการตรวจสอบนั้นควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ได้

2. พิจารณาความน่าจะเป็นซึ่งค่าผิดปรกติที่ถูกตรวจพบเป็นค่าผิดปรกติจริงทุกค่า (p_1), ความน่าจะเป็นซึ่งทำให้เกิดมาซคคิงเอฟเฟ็ค (p_2) และความน่าจะเป็นซึ่งทำให้เกิดชวอมพิงเอฟเฟ็ค (p_3) เราจะทำการเปรียบเทียบเฉพาะตัวสถิติทดสอบที่สามารถควบคุมของความน่าจะเป็นความผิดพลาดประเภทที่ 1 ได้เท่านั้น

1.8 ประโยชน์ที่คาดว่าจะได้รับ

ผลการศึกษาทำให้ผู้ใช้สามารถเลือกตัวสถิติทดสอบที่เหมาะสมในกรณีที่ข้อมูลมีค่าผิดปรกติหลายค่าในสมการการถดถอยเชิงเส้น