การตรวจสอบค่าผิดปกติตามแนวทางของเบย์ :
ตัวสถิติระยะทางคูลแบคก์และ L_1 ในรูปอย่างง่าย

โดย
นายประสพชัย พสุนนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์
ภาควิชาคณิตศาสตร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2545

ISBN 974 - 653 - 194 - 8

A BAYESIAN APPROACH TO OUTLIER DETECTION : A SIMPLIFIED VERSION OF KULLBACK AND L_1 DISTANCE STATISTICS

By

Prasopchai Phasunon

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Department of Mathematics

Graduate School

SILPAKORN UNIVERSITY

2002

ISBN 974 - 653 - 194 - 8

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้วิทยานิพนธ์เรื่อง " การตรวจสอบค่าผิดปกติตามแนวทางของเบย์ : ตัวสถิติระยะทางคูลแบคก์ และ L_1 ในรูปอย่างง่าย " เสนอโดย นายประสพชัย พสุนนท์ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์

.....

(ผู้ช่วยศาสตราจารย์ ดร.จิราวรรณ คงคล้าย)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... เดือน ..... พ.ศ. .....

ผู้ควบคุมวิทยานิพนธ์

1. ผู้ช่วยศาสตราจารย์ ดร.ปราณี นิลกรณ์
2. รองศาสตราจารย์ ดร.สุดา ตระการเถลิงศักดิ์

คณะกรรมการตรวจสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ วีรานันท์ พงศาภักดี)
..... / ..... / .....

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ปราณี นิลกรณ์)
..... / ..... / .....

..... กรรมการ
(รองศาสตราจารย์ ดร.สุดา ตระการเถลิงศักดิ์)
..... / ..... / .....

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.กมล บุษบา)
..... / ..... / .....

..... กรรมการ
(รองศาสตราจารย์ ไพบูลย์ รัตนประเสริฐ)
..... / ..... / .....

K 42515003 : สาขาวิชาสถิติประยุกต์
คำสำคัญ : ค่าผิดปกติ / ความอ่อนไหวแบบเบย์ / ระยะทางคูลแบคก์ / ระยะทาง L_1

ประสพชัย พสุนนท์ : การตรวจสอบค่าผิดปกติตามแนวทางของเบย์ : ตัวสถิติระยะทางคูลแบคก์ และ L_1 ในรูปอย่างง่าย (A BAYESIAN APPROACH TO OUTLIER DETECTION : A SIMPLIFIED VERSION OF KULLBACK AND L_1 DISTANCE STATISTICS) อาจารย์ผู้ควบคุมวิทยานิพนธ์ : ผศ. ดร.ปราณี นิลกรณ์ และ รศ. ดร.สุดา ตระการเถลิงศักดิ์. 113 หน้า. ISBN 974 - 653 - 194 - 8

การวิจัยนี้มีวัตถุประสงค์ เพื่อ (1) พัฒนาตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติตามแนวทางของเบย์ให้สามารถใช้ตรวจสอบค่าผิดปกติได้ง่ายขึ้น โดยการประมาณตัวสถิติระยะทางคูลแบคก์ด้วยตัวสถิติ $\tilde{K}$ และประมาณตัวสถิติระยะทาง L_1 ด้วยตัวสถิติ $\tilde{L}_1$ และ $\acute{L}_1$ ทั้ง 3 ตัวสถิติที่พัฒนาขึ้นมีพื้นฐานจากวิธีการแบบเบย์ที่ใช้การแจกแจงก่อนแบบไม่มีสารสนเทศ (2) เปรียบเทียบผลการตรวจสอบค่าผิดปกติของตัวสถิติทั้ง 3 และของตัวสถิติ Generalized Extreme Studentized Deviate (GESD) ซึ่งใช้แนวทางแบบดั้งเดิมในการตรวจสอบค่าผิดปกติ ข้อมูลที่ใช้ในการตรวจสอบค่าผิดปกติได้จากการจำลองแบบและชุดข้อมูลจริง ข้อมูลจากการจำลองแบบสุ่มจากประชากรที่มีการแจกแจงแบบปกติมาตรฐานโดยใช้ขนาดตัวอย่าง 10 , 20 , 50 และ 80 และปะปนค่าผิดปกติ 1 ค่า โดยศึกษาค่าผิดปกติ 3 ขนาด คือ ขนาดเล็ก ($3\sigma^2$) ขนาดกลาง ($4\sigma^2$) และขนาดใหญ่ ($6\sigma^2$) จากนั้นใช้ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$, $\acute{L}_1$ และ GESD ตรวจสอบค่าผิดปกติในข้อมูลตัวอย่างที่จำลองแบบขึ้น โดยทำซ้ำ 2,000 ครั้ง

ผลการวิจัยพบว่า

1. สำหรับข้อมูลตัวอย่างที่จำลองแบบ กรณีตัวอย่างขนาด 10 พบว่า สำหรับค่าผิดปกติขนาดเล็กและขนาดกลาง ตัวสถิติ $\tilde{K}$ มีจำนวนการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ $\acute{L}_1$, $\tilde{L}_1$ และ GESD ตามลำดับ ส่วนค่าผิดปกติขนาดใหญ่พบว่าตัวสถิติ GESD มีจำนวนการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ $\tilde{K}$, $\acute{L}_1$ และ $\tilde{L}_1$ ตามลำดับ กรณีขนาดตัวอย่าง 20 , 50 และ 80 พบว่า สำหรับค่าผิดปกติขนาดเล็ก ตัวสถิติ $\tilde{K}$ มีจำนวนการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ $\acute{L}_1$, $\tilde{L}_1$ และ GESD ตามลำดับ ส่วนค่าผิดปกติขนาดกลางและขนาดใหญ่พบว่าตัวสถิติ GESD มีจำนวนการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ $\tilde{K}$, $\acute{L}_1$ และ $\tilde{L}_1$ ตามลำดับ

2. สำหรับชุดข้อมูลจริง ทำการศึกษาข้อมูล 3 ชุด คือ ข้อมูลของ Freeman พบว่าตัวสถิติ $\tilde{K}$,$\tilde{L}_1$ และ $\acute{L}_1$ ตรวจพบค่าผิดปกติ 2 ค่า ส่วนตัวสถิติ GESD ตรวจพบค่าผิดปกติ 1 ค่า ข้อมูลของ Darwin พบว่าตัวสถิติ $\tilde{K}$, $\tilde{L}_1$, $\acute{L}_1$ และ GESD ตรวจพบค่าผิดปกติ 2 ค่า และข้อมูลของ Sacks et al. พบว่าตัวสถิติ $\tilde{K}$, $\tilde{L}_1$, $\acute{L}_1$ และ GESD ตรวจพบค่าผิดปกติ 3 ค่า

ภาควิชาคณิตศาสตร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2545

ลายมือชื่อนักศึกษา .....

ลายมือชื่ออาจารย์ผู้ควบคุมวิทยานิพนธ์ 1. ..... 2. .....

K 42515003 : MAJOR : APPLIED STATISTICS

KEY WORD : OUTLIERS / BEYESIAN SENSITIVITY / KULLBACK DISTANCE / L_1 DISTANCE

PRASOPCHAI PHASUNON : A BAYESIAN APPROACH TO OUTLIER DETECTION : A SIMPLIFIED VERSION OF KULLBACK AND L_1 DISTANCE STATISTICS. THESIS ADVISORS : ASST. PROF. PRANEE NILKORN , Ph.D., AND ASSO. PROF. SUDA TRAKANTHARAUGSAK , Ph.D.

113 pp. ISBN 974 - 653 - 194 - 8

The objectives of this research are (1) to develop a simplified version of outlier detection statistics based on Kullback and L_1 distances under Bayesian approach with noninformative prior. The statistics developed are $\tilde{K}$, an approximation of Kullback distance , and $\tilde{L}_1$ and $\tilde{L}_1'$, approximations of L_1 distance. (2) to compare the behavior of the developed statistics and of Generalized Extreme Studentized Deviate (GESD) statistic based on the classical approach. Both simulated and real data are used. For simulated data , 2,000 samples are generated from a standard normal population and an outlier is contaminated for each sample. The simulations are repeated for 4 different sample sizes , 10 , 20 , 50 and 80 , and 3 different magnitudes of outliers , small size $(3\sigma^2)$, medium size $(4\sigma^2)$ and large size $(6\sigma^2)$.

The results of the study indicate that :

1. For simulated samples of size 10 , with small and medium magnitudes of contaminated outliers , $\tilde{K}$ is able to detect outers most frequently , followed by $\tilde{L}_1'$, $\tilde{L}_1$ and GESD respectively. With large magnitude of outliers , GESD can detect outlier most frequently , followed by $\tilde{K}$, $\tilde{L}_1'$ and $\tilde{L}_1$ respectively. For samples of size 20 , 50 , and 80 , with small magnitude of outlier , $\tilde{K}$ is able to detect outers most frequently , followed by $\tilde{L}_1'$, $\tilde{L}_1$ and GESD respectively. With medium and large magnitudes of outliers , GESD can detect outlier most frequently , followed by $\tilde{K}$, $\tilde{L}_1'$ and $\tilde{L}_1$ respectively.

2. Three real data sets are studied. With Freeman's data , two outliers are identified by $\tilde{K}$, $\tilde{L}_1$ and $\tilde{L}_1'$ while only one outlier is identified by GESD. With Dawin's data , 2 outliers are identified by all 4 statistics. With Sacks et al.'s data , 3 outliers are identified by all 4 statistics.

Department of Mathematics Graduate School , Silpakorn Univesity Academic Year 2002

Student's signature ……

Thesis Advisors' signature 1. ……. 2. …….

กิตติกรรมประกาศ

เป็นที่ยอมรับกันว่า ทักษะการเขียนเป็นทักษะที่ยากที่สุดทักษะหนึ่ง โดยเฉพาะงานเขียนที่ต้องอาศัยความรู้ทางวิชาการขั้นสูง วิทยานิพนธ์ฉบับนี้ ถือเป็นปฐมบทแรกในเชิงวิชาการของผู้วิจัย ซึ่งยังอ่อนประสบการณ์นัก การวิจัยนี้สำเร็จสมบูรณ์ได้ ผู้วิจัยกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.ปราณี นิลกรณ์ อาจารย์ผู้เมตตา สั่งสอนผู้วิจัยด้วยความรักเสมอมาแม้ศิษย์นี้เขลาเพียงใด และ รองศาสตราจารย์ ดร.สุดา ตระการเถลิงศักดิ์ สำหรับข้อเสนอแนะอันเป็นประโยชน์ยิ่งต่อผู้วิจัย

กราบขอบพระคุณ รองศาสตราจารย์ วีรานันท์ พงศาภักดี ประธานคณะกรรมการตรวจสอบวิทยานิพนธ์ สำหรับแนวคิดที่มีคุณค่าต่อผู้วิจัยทั้งในเรื่องเรียนและการใช้ชีวิต กราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.กมล บุษบา และรองศาสตราจารย์ ไพบูลย์ รัตนประเสริฐที่เสียสละเวลาในการอ่าน ซักถาม และคำแนะนำอันมีค่ายิ่งสำหรับการวิจัยครั้งนี้ นอกจากนี้ ผู้วิจัยขอกราบขอบพระคุณ ครู อาจารย์ทุกท่านที่ประสิทธิ์ประสาทความรู้ต่อผู้วิจัย ศิษย์ได้ดีเพราะมีครู

ขอขอบคุณ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร และทบวงมหาวิทยาลัย สำหรับทุนอุดหนุนในการทำวิจัยครั้งนี้

ขอขอบคุณ คณาจารย์ สถาบันราชภัฏนครปฐม โดยเฉพาะพี่และเพื่อนอาจารย์ในโปรแกรมวิชาคณิตศาสตร์และสถิติประยุกต์ สำหรับความมีน้ำใจ และความปรารถนาดีที่มีให้

ขอบคุณ คุณสิริโชค พสุนนท์ สำหรับความห่วงใยในพี่ชายเสมอมา และขอบคุณ คุณอัจฉรา โกษาแสง กัลยาณมิตรรุ่นพี่ สำหรับไมตรีจิตที่มอบให้

ท้ายสุด ผู้วิจัยกราบขอบพระคุณ คุณพ่อ คุณแม่ สำหรับความรัก กำลังใจที่ไม่เคยขาดหาย ไม่ว่าในเวลาที่สุขหรือทุกข์ ท่านพร้อมที่จะให้เสมอ เป็นความรักที่ยิ่งใหญ่เกินจะอธิบายได้

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในยุคปัจจุบัน ข้อมูลสารสนเทศ (Information data) มีความสำคัญมาก จำเป็นที่ต้องใช้ข้อมูลที่มีคุณภาพ เพื่อนำไปวิเคราะห์วางแผนหรือตัดสินใจให้เกิดประโยชน์สูงสุด ดังนั้นข้อมูลที่ใช้ควรมีความถูกต้องเชื่อถือได้ บ่อยครั้งที่พบว่าข้อมูลที่ได้จากการเก็บรวบรวมหรือจากการทดลองมีค่าของข้อมูลบางค่าแตกต่างไปจากข้อมูลส่วนใหญ่ กล่าวคือ อาจมีบางค่าที่มีค่าสูงเกินไปหรือต่ำเกินไปเมื่อเทียบกับข้อมูลตัวอื่น ๆ เราเรียกค่าเหล่านั้นว่า **ค่าผิดปกติ** (Outlier) ค่าผิดปกติที่เกิดขึ้นสามารถพิจารณาได้ 2 ลักษณะ คือ

1. ค่าผิดปกติที่เกิดขึ้นเป็นค่าที่มีความน่าสนใจในตัวเอง เช่น เป็นค่าที่แทนปริมาณแร่ธาตุที่มีนัยสำคัญต่อการพัฒนาผลิตผลการเกษตร

2. ในตัวอย่างหากเกิดค่าผิดปกติขึ้น อาจมีสาเหตุมาจากความคลาดเคลื่อนในข้อมูล ซึ่งความคลาดเคลื่อนแบ่งเป็น 3 ประเภท ดังนี้

2.1 ความคลาดเคลื่อนจากความแปรผันที่มีในประชากรที่ทำการศึกษา (Inherent variability) ไม่ได้เกิดจากการวัดหรือการเก็บข้อมูลที่ผิดพลาด เป็นความคลาดเคลื่อนที่หลีกเลี่ยงไม่ได้ เพราะไม่ว่าจะควบคุมการวัดหรือเก็บข้อมูลให้ดีเพียงใด ความแปรผันในประชากรก็ยังคงอยู่

2.2 ความคลาดเคลื่อนจากการวัด (Measurement error) เป็นความคลาดเคลื่อนที่เกิดจากการใช้เครื่องมือในการวัดที่มีคุณภาพต่ำ

2.3 ความคลาดเคลื่อนจากการปฏิบัติการ (Execution error) เป็นความคลาดเคลื่อนที่เกิดขึ้นจากความผิดพลาดของการบันทึกข้อมูลเพื่อประมวลผล เช่น การลงรหัส หรืออาจเกิดจากการเลือกตัวอย่างที่ไม่เหมาะสมกับประชากร

กรณีที่ค่าผิดปกติเป็นค่าที่มีความน่าสนใจในตัวเอง ถ้าตัดข้อมูลที่มีความผิดปกติออกก่อนที่จะทำการวิเคราะห์ข้อมูลอาจทำให้ขาดสารสนเทศบางอย่างไป แต่ถ้ากรณีที่ค่าผิดปกติเกิดขึ้นเนื่องจากความคลาดเคลื่อนในการวัดหรือเกิดจากการปฏิบัติการที่มีความคลาดเคลื่อน อาจจำเป็นต้องตัดข้อมูลที่มีความผิดปกติออกก่อนที่จะวิเคราะห์เพื่อให้ผลที่ได้จากการวิเคราะห์

เชื่อถือได้ การตรวจสอบค่าผิดปกติจึงถือเป็นเรื่องสำคัญที่ควรทำก่อนที่จะทำการลงมือวิเคราะห์ข้อมูล ซึ่งจะทำให้ผู้วิเคราะห์ข้อมูลเห็นลักษณะโครงสร้างข้อมูล และอาจหาวิธีการแก้ปัญหาด้วยวิธีการอย่างหนึ่งอย่างใดได้เหมาะสม

ในการวิเคราะห์ข้อมูลทางสถิติจากตัวอย่างสุ่ม (Random sample) ที่มีข้อสมมติ (Assumption) ว่าประชากรมีการแจกแจงแบบปกติ นั่นคือ ประชากรมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ μ และความแปรปรวนเท่ากับ σ^2 ในความเป็นจริงข้อมูลที่ได้อาจมีการแจกแจงเบี่ยงเบนไปจากข้อสมมติ กล่าวคือ มีการแจกแจงที่ไม่เป็นปกติ ได้แก่ การแจกแจงแบบปลอมปน (Contaminated distribution) การแจกแจงที่มีหางยาว (Long tails) หรือการแจกแจงที่มีหางหนา (Heavy tails) ลักษณะดังกล่าวส่งผลต่อการวิเคราะห์ข้อมูล ผลการวิเคราะห์อาจไม่ถูกต้อง วิธีการทางสถิติ เช่น การวิเคราะห์การถดถอย (Regression analysis) การวิเคราะห์ตัวประกอบ (Factor analysis) การวิเคราะห์ความแปรปรวน (ANOVA) ฯลฯ ล้วนมีข้อสมมติว่าข้อมูลตัวอย่างสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ ดังนั้น ก่อนการวิเคราะห์ข้อมูลทางสถิติสำหรับข้อมูลตัวอย่างที่สุ่มจากประชากร ผู้วิเคราะห์ควรทำการตรวจสอบข้อมูลเบื้องต้นว่าเป็นไปตามข้อสมมติหรือไม่ หากพบค่าผิดปกติจะได้แก้ไขปัญหาก่อนการวิเคราะห์ข้อมูล เพื่อให้ได้ผลการวิเคราะห์ที่ถูกต้อง การตรวจสอบค่าผิดปกติมีอยู่หลายวิธี การใช้กราฟเป็นวิธีหนึ่งที่ใช้ได้ดี แต่เนื่องจากการใช้กราฟเป็นวิธีที่ให้ผลสรุปที่ไม่แน่นอนขึ้นอยู่กับความเชี่ยวชาญและประสบการณ์ของผู้วิเคราะห์ ดังนั้น การที่จะตัดสินใจว่าค่าที่สงสัยนั้นเป็นค่าผิดปกติหรือไม่ ควรจะมีการตรวจสอบอย่างมีกฎเกณฑ์

ปัญหาค่าผิดปกติในข้อมูลที่มาจากการแจกแจงแบบปกติอาจอธิบายโดยใช้สัญลักษณ์ดังนี้ ให้ $\underline{x} = \{x_1, x_2, ..., x_n\}$ เป็นเซตข้อมูลของตัวอย่างที่ประกอบด้วยสับเซต 2 สับเซตคือ $J_1 = \{ x_{11}, x_{12}, ..., x_{1m} \}$ และ $J_2 = \{ x_{21}, x_{22}, ..., x_{2k} \}$ โดยที่ $X_{1p} \sim N(\mu, \sigma^2)$, p = 1,2,...,m และ $X_{2q} \sim N(\mu_q, \sigma^2_q)$, q = 1,2, ... ,k นั่นคือ สมาชิกในเซต J_1 มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยและความแปรปรวนคงที่ ส่วนสมาชิกในเซต J_2 มีการแจกแจงแบบปกติแต่ค่าเฉลี่ยและความแปรปรวนไม่คงที่ เราเรียก J_1 ว่าเป็นเซตข้อมูลปกติ (Inlier data set) และเรียก J_2 ว่าเป็นเซตข้อมูลผิดปกติ (Outlier data set) โดยที่ $m+k \equiv n$

ในการอนุมานทางสถิติแบบเบย์ ต้องอาศัยสารสนเทศ 2 ส่วน คือ (1) ความรู้ที่มีอยู่ก่อนเกี่ยวกับพารามิเตอร์ที่สนใจ ซึ่งแทนความรู้ดังกล่าวในรูปของการแจกแจงก่อน (Prior distribution) และ (2) ความรู้ที่ได้จากข้อมูลที่รวบรวมมาซึ่งแทนด้วยภาวะน่าจะเป็น (Likelihood) ดังนั้น ปัญหาค่าผิดปกติในชุดข้อมูลจึงมีผลต่อการอนุมานแบบเบย์เช่นกันโดยส่งผลผ่านทางภาวะน่าจะเป็น ถ้าชุดข้อมูลมีค่าผิดปกติผสมมาด้วย ผลจากค่าผิดปกติจะทำให้การแจกแจงภายหลัง (Posterior distribution) เบี่ยงเบนไปจากที่ควรจะเป็น ในทางสถิติ มีผู้ศึกษาความอ่อนไหว (Sensitivity) ของการอนุมานแบบเบย์หลายลักษณะด้วยกัน เช่น ความอ่อนไหวของการอนุมานอันเนื่องมาจากการกำหนดการแจกแจงก่อนผิดพลาด (Prior misspecification) ความอ่อนไหวเนื่องจากข้อมูลมีค่าผิดปกติ ความอ่อนไหวเนื่องจากเลือกใช้ฟังก์ชันภาวะน่าจะเป็นไม่ถูกต้อง (Misspecification of likelihood function) เป็นต้น

Box and Tiao (1973) ตรวจสอบอิทธิพลของความไม่เป็นปกติ (Effect of non-normality) ของข้อมูลที่มาจากประชากรที่มีการแจกแจงแบบปกติ ตามแนวทางของ Karl Pearson โดยตัวสถิติที่ใช้วัดความไม่เป็นปกติ คือ (1) ตัวสถิติ $\gamma_1 = \frac{E(y-\mu)^3}{\sigma^3}$ ซึ่งใช้ประเมินความเบ้ (Skewness) ถ้า $\gamma_1 < 0$ โค้งจะเบ้ซ้าย $\gamma_1 > 0$ โค้งจะเบ้ขวา และ $\gamma_1 = 0$ โค้งสมมาตร และ (2) ตัวสถิติ $\gamma_2 = \frac{E(y-\mu)^4}{\sigma^4} - 3$ ซึ่งใช้ประเมินความโด่ง (Kurtosis) ถ้า $\gamma_2 < 0$ โค้งจะโด่งน้อย $\gamma_2 > 0$ โค้งจะโด่งมาก และ $\gamma_2 = 0$ โค้งจะโด่งปานกลาง

Geisser (1980) เสนอตัวสถิติ Conditional Predictive Ordinate (CPO) ในการตรวจสอบความอ่อนไหวเนื่องจากข้อมูลปะปนค่าผิดปกติ โดยตัวสถิติ CPO เป็นตัวสถิติที่ได้จากการวิเคราะห์โดยใช้หลักการสถิติแบบเบย์เป็นพื้นฐาน (รายละเอียดแสดงไว้หน้า 30 , 33)

Geisser (1989) ใช้ตัวสถิติตามแนวทางเบย์ ในการวัดความไม่ลงรอยกัน (Discordancy) ของค่าสังเกตที่แจกแจงแบบเอกซ์โพเนนเชียล ตัวสถิติดังกล่าวคือ ตัวสถิติ Conditional Predictive Discordancy (CPD) และตัวสถิติ Unconditional Predictive Discordancy (UPD) โดยค่าสังเกตที่ถูกวัดแล้วไม่ลงรอยกันแสดงว่ามีความอ่อนไหวระหว่างตัวแบบ 2 ตัวแบบ คือ ตัวแบบที่ใช้ค่าสังเกตครบทุกค่ากับตัวแบบที่ตัดค่าสังเกตออก 1 ค่า ลักษณะของการอ่อนไหวเช่นนี้น่าจะมีสาเหตุเกิดจากข้อมูลมีค่าผิดปกติปะปน

Kass , Tierney and Kadane (1989) เสนอวิธีประมาณค่าคาดหวังของการแจกแจงภายหลัง (Posterior expectation) โดยวิธีการกระจายแบบ Laplace (Laplace expansion) สำหรับใช้ตรวจสอบความอ่อนไหวอันเนื่องมาจากการกำหนดการแจกแจงก่อนผิดพลาด และใช้ตรวจสอบความอ่อนไหวอันเนื่องมาจากมีค่าผิดปกติในชุดข้อมูล โดยตัวสถิติที่ได้จากการประมาณโดยวิธีการกระจายแบบ Laplace คือ ตัวสถิติที่ใช้วัดการเปลี่ยนแปลงที่มีค่ามาตรฐานสูงสุด (Maximum Standardized Change Statistic) ซึ่งสามารถตรวจสอบความอ่อนไหวทั้งสองลักษณะโดยผ่านฟังก์ชันการก่อกวน (Perturbation function)

Pettit (1990) ศึกษาความอ่อนไหวเนื่องจากข้อมูลมีค่าผิดปกติ โดยเปรียบเทียบตัวสถิติ CPO และตัวสถิติ Ratio Ordinate Measure (ROM) ในข้อมูลที่มีการแจกแจงแบบปกติในตัวแปรเดียวและหลายตัวแปร ทั้งที่ทราบความแปรปรวนและไม่ทราบความแปรปรวนของประชากร โดยตัวสถิติ ROM พัฒนามาจากตัวสถิติ CPO (รายละเอียดแสดงไว้หน้า 31)

Pettit (1992) ใช้ตัวประกอบเบย์ (Bayes factor) ในการตรวจสอบความอ่อนไหวเนื่องจากค่าผิดปกติ โดยตรวจสอบตัวแบบที่มีค่าผิดปกติปะปนมาในชุดข้อมูลที่มาจากการแจกแจงแบบปกติ ข้อมูลที่มีการแจกแจงแบบเอกซ์โพเนนเชียล และการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (รายละเอียดแสดงไว้หน้า 32)

Geisser (1993) ให้การทดสอบนัยสำคัญการพยากรณ์ สำหรับความไม่ลงรอย (Predictive signficance testing for discordancy) ในการวัดความอ่อนไหวเนื่องจากข้อมูลมีค่าผิดปกติ และความอ่อนไวเนื่องจากใช้ฟังก์ชันภาวะน่าจะเป็นไม่ถูกต้อง (รายละเอียดแสดงไว้หน้า 33)

Weiss (1996) ประเมินความอ่อนไหว อันเนื่องมาจากข้อมูลมีค่าผิดปกติปะปนในการอนุมานแบบเบย์ ผ่านความแตกต่างระหว่างข้อมูลสมบูรณ์ที่ใช้ค่าสังเกตครบทุกค่าและข้อมูลที่ตัดค่าสังเกตออก 1 ค่า โดยใช้การวัดการเบี่ยงเบนของตัวสถิติระยะทาง L_1 และตัวสถิติระยะทาง χ^2 ตรวจสอบค่าผิดปกติที่ละค่าสังเกต

ในกรณีที่นักวิจัยต้องการวิเคราะห์ข้อมูลโดยใช้วิธีอนุมานแบบเบย์ แต่นักวิจัยไม่มีความรู้ก่อนเกี่ยวกับพารามิเตอร์นั้น ในทางปฏิบัตินิยมใช้การแจกแจงก่อนแบบไม่มีสารสนเทศ (Noninformative prior) ซึ่งในกรณีเช่นนี้ ไม่จำเป็นต้องพิจารณาความอ่อนไหวของการอนุมานอันเนื่องมาจากการแจกแจงก่อน ความอ่อนไหวของการอนุมานแบบเบย์จะขึ้นอยู่กับการเลือกใช้ฟังก์ชันภาวะน่าจะเป็นและความถูกต้องของข้อมูล ดังกล่าวแล้วว่าการวิเคราะห์ข้อมูลส่วนใหญ่ล้วนมีข้อสมมติว่าข้อมูลมาจากประชากรที่มีการแจกแจงแบบปกติ ดังนั้นถ้าถือว่าข้อสมมติถูกต้อง ความอ่อนไหวของการอนุมานจะขึ้นอยู่กับข้อมูลเท่านั้น การตรวจสอบค่าผิดปกติในการอนุมานแบบเบย์จึงเป็นเรื่องสำคัญเช่นกัน

การตรวจสอบค่าผิดปกติมีวิธีการที่เสนอไว้หลายวิธี ทั้งนี้การเลือกวิธีการหรือตัวสถิติในการตรวจสอบค่าผิดปกติขึ้นอยู่กับความเชี่ยวชาญและวัตถุประสงค์ของนักสถิติแต่ละท่าน แนวทางแรกเป็นการตรวจสอบค่าผิดปกติซึ่งเป็นแนวคิดตามระเบียบวิธีทางสถิติแบบดั้งเดิม (Classical statistics) และแนวทางที่สองเป็นแนวคิดตามระเบียบวิธีทางสถิติแบบเบย์ (Bayesian statistics) มีรายละเอียดดังนี้

1. การตรวจสอบค่าผิดปกติตามระเบียบวิธีทางสถิติแบบดั้งเดิมซึ่งมีหลายวิธีด้วยกัน เช่น วิธีของ Grubbs (1950) วิธีของ Dixon (1953) วิธีของ Ferguson (1961) วิธีของ Tietjen and Moore (1972) วิธีของ Tukey (1977) วิธีของ Rosner (1983) วิธีการที่น่าสนใจวิธีหนึ่ง คือ การตรวจสอบค่าผิดปกติโดยวิธีการ Generalized Extreme Studentized Deviate (GESD) ของ Rosner (1983) ซึ่งใช้ตรวจสอบค่าผิดปกติได้ 1 ถึง k ค่าในหนึ่งชุดข้อมูล ตัวสถิติ GESD จะพิจารณาอัตราส่วนของระยะทางมากที่สุดระหว่างค่าสังเกตแต่ละค่ากับค่าเฉลี่ยต่อส่วนเบี่ยงเบนมาตรฐาน โดยในการตรวจสอบค่าผิดปกติของตัวสถิติ GESD นี้จะใช้วิธีการตรวจสอบค่าผิดปกติแบบหลายค่า (Many outliers procedure) เนื่องจากวิธีนี้มีความสามารถในการตรวจสอบค่าผิดปกติได้ตั้งแต่ 1 ถึง k ค่า จากข้อมูลขนาด n โดยการสมมติก่อนการตรวจสอบค่าผิดปกติว่าชุดข้อมูลที่เราตรวจสอบค่าผิดปกตินั้น มีค่าผิดปกติปะปนอยู่ k ค่าโดยที่ k ควรจะน้อยกว่า $\frac{n}{2}$

2. วิธีการวิเคราะห์ความอ่อนไหวแบบเบย์ (Bayesian sensitivity analysis) ตัวสถิติที่ใช้ในวิธีนี้เป็นตัวสถิติที่วัดความอ่อนไหวระหว่างตัวแบบ 2 ตัวแบบ โดยตัวแบบแรกเป็นตัวแบบสมบูรณ์ที่ใช้ค่าสังเกตครบทุกค่าในชุดข้อมูล ส่วนอีกตัวแบบหนึ่งเป็นตัวแบบที่ถูกตัดค่าสังเกตตัวที่ i ออกจากชุดข้อมูล วิธีการประเมินค่าความผิดปกติในชุดข้อมูลจะประเมินจากค่าสถิติที่ได้จากการคำนวณระยะทางระหว่าง 2 ตัวแบบ โดยถ้าค่าสถิติที่ได้มีค่ามากแสดงว่ามี

ความแตกต่างระหว่าง 2 ตัวแบบมาก ซึ่งเป็นอิทธิพลของค่าสังเกตตัวที่ i ที่ถูกตัดออกไปจากชุดข้อมูล จะถือว่าค่าสังเกตตัวที่ i นั้นเป็นค่าผิดปกติในข้อมูลชุดนั้น ในทางตรงกันข้ามหากค่าสถิติที่ได้จากวิธีการวิเคราะห์ความอ่อนไหวแบบเบย์มีค่าน้อยแสดงว่าตัวแบบทั้ง 2 ตัวแบบมีความแตกต่างกันน้อย ดังนั้น ค่าสังเกตตัวที่ i ที่ถูกตัดค่าออกไม่ได้มีอิทธิพลมากนักต่อตัวแบบสมบูรณ์ จะถือว่าค่าสังเกตตัวที่ i นั้นเป็นค่าปกติในชุดข้อมูล วิธีการวิเคราะห์ความอ่อนไหวแบบเบย์มีพื้นฐานจากตัวประกอบเบย์ โดยวิธีการตรวจสอบค่าผิดปกติจะใช้วิธีการตรวจสอบค่าผิดปกติทีละ 1 ค่า (One outlier procedure) ตัวสถิติที่นำมาใช้ในวิธีการวิเคราะห์ความอ่อนไหวแบบเบย์มีหลายตัวด้วยกัน เช่น ตัวสถิติคูลแบคก์ (Kullback statistic) ตัวสถิติเบอนาร์โด (Bernardo statistic) ตัวสถิติระยะทาง L_1 (L_1 - distance statistic) ตัวสถิติที่ใช้วัดการเปลี่ยนแปลงที่มีค่ามาตรฐานสูงสุด ตัวสถิติวัดความเบี่ยงเบนแบบ χ^2 (χ^2 - divergence statistic) เป็นต้น ปัญหาสำคัญของตัวสถิติที่ใช้วิธีการวิเคราะห์ความอ่อนไหวแบบเบย์ คือ ไม่มีเกณฑ์ที่แน่นอนในการประเมินค่าความผิดปกติของข้อมูล จากการศึกษาของ Weiss (1996) มีตัวสถิติระยะทาง 2 ตัวที่น่าสนใจและมีพื้นฐานจากตัวประกอบแบบเบย์ คือ

2.1 ตัวสถิติที่วัดการเบี่ยงเบนของคูลแบคก์ (Kullback divergence statistic) หรือตัวสถิติคูลแบคก์ (Kullback statistic) การตรวจสอบค่าผิดปกติโดยใช้การวัดการเบี่ยงเบนของคูลแบคก์ มีพื้นฐานจากการมองความแตกต่างของตัวแบบสมบูรณ์และตัวแบบที่ตัดค่าสังเกตตัวที่ i ออก ตัวสถิติคูลแบคก์เป็นตัวสถิติที่วัดระยะทาง โดยระยะทางดังกล่าว คือ ค่าของลอการิทึมธรรมชาติ (Nature logarithm) ของอัตราส่วนระหว่างตัวแบบสมบูรณ์ที่ใช้ค่าสังเกตครบทุกค่ากับตัวแบบที่ตัดค่าสังเกตตัวที่ i ออก โดยทั้งสองตัวแบบเป็นการแจกแจงภายหลัง (Posterior distribution) สูตรทั่วไปของสถิติคูลแบคก์ คือ $\int -\log_e \left[\frac{p_1\left(\theta \middle| x^i\right)}{p\left(\theta \middle| \underline{x}\right)} \right] p\left(\theta \middle| \underline{x}\right) d\theta$

เมื่อ $p_1\left(\theta \middle| x^i\right)$ คือ ฟังก์ชันความน่าจะเป็นของการแจกแจงภายหลังของ θ เมื่อตัดค่าสังเกตตัวที่ i ออก (1.1)

$p\left(\theta \middle| \underline{x}\right)$ คือ ฟังก์ชันความน่าจะเป็นของการแจกแจงภายหลังของ θ เมื่อใช้ค่าสังเกตครบทุกค่า (1.2)

2.2 ตัวสถิติที่ใช้วัดอิทธิพลแบบระยะทาง L_1 (L_1 - distance influence Statistic) หรือตัวสถิติระยะทาง L_1 (L_1 - distance statistic) ซึ่งใช้วัดระยะห่างระหว่างตัวแบบสมบูรณ์ที่ใช้ค่าสังเกตครบทุกค่ากับตัวแบบที่ตัดค่าสังเกตออก 1 ค่า เพื่อประเมินความอ่อนไหวของค่าสังเกตแต่ละค่าจนครบทุกค่าสังเกต โดยดูที่ระยะทางระหว่างการแจกแจงภายหลัง ภายใต้ 2 ตัวแบบ คือ ตัวแบบสมบูรณ์และตัวแบบที่ถูกตัดค่าสังเกต i ออก เช่นเดียวกับตัวสถิติคูลแบคก์ ตัวสถิติระยะทาง L_1 มีสูตรทั่วไป คือ $0.5\int \left| p_1\left(\theta \middle| x^i\right) - p\left(\theta \middle| \underline{x}\right)\right| d\theta$

เมื่อ $p_1\left(\theta \middle| x^i\right)$ และ $p\left(\theta \middle| \underline{x}\right)$ ดังแสดงใน (1.1) และ (1.2) ตามลำดับ

ดังที่กล่าวมาแล้วข้างต้นว่า ข้อมูลที่มาจากการแจกแจงแบบปกตินั้นมีความสำคัญ การที่ข้อมูลมีค่าผิดปกติมาปะปนในข้อมูลที่ส่วนใหญ่มาจากการแจกแจงแบบปกติ จะทำให้การวิเคราะห์ข้อมูลเกิดความผิดพลาดหรือคลาดเคลื่อนได้ สำหรับการตรวจสอบค่าผิดปกติโดยวิธีสถิติแบบดั้งเดิม Rosner (1983) ได้เสนอตัวสถิติ GESD ที่ใช้ตรวจสอบค่าผิดปกติแบบหลายค่าในชุดข้อมูล โดยที่ตัวสถิติ GESD นี้สามารถตรวจสอบค่าผิดปกติได้อย่างมีประสิทธิภาพในข้อมูลที่มีขนาดมากกว่า 25 หากข้อมูลที่นำมาตรวจสอบค่าผิดปกติมีขนาดน้อยกว่า 25 ความน่าเชื่อถือในตัวสถิติ GESD นี้ยังน่าสงสัยอยู่สำหรับค่าผิดปกติที่ตรวจพบ ในการตรวจสอบค่าผิดปกติแบบเบย์ Weiss (1996) ได้ใช้การตรวจโดยระยะทางที่ได้จากการวัดการเบี่ยงเบน (Divergence Measure) ระหว่างตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวน โดยมีตัวประกอบเบย์เป็นพื้นฐานในการประเมินค่าความผิดปกติของข้อมูล โดยที่ตัวแบบสมบูรณ์เป็นการแจกแจงภายหลังเมื่อใช้ค่าสังเกตครบทุกค่า และตัวแบบที่ถูกก่อกวนเป็นการแจกแจงภายหลังเมื่อตัดค่าสังเกตตัวที่ i ออก

ตัวสถิติที่ Weiss (1996) ใช้ในการตรวจสอบค่าผิดปกติ คือ ตัวสถิติ CPO ตัวสถิติระยะทาง L_1 ตัวสถิติเบอร์นาโด และตัวสถิติ χ^2 ซึ่งตัวสถิติทั้งหมดให้ผลลัพธ์ที่สอดคล้องกันในการตรวจสอบค่าผิดปกติ เป็นที่น่าสังเกตว่าตัวสถิติระยะทาง L_1 นั้นสามารถประเมินค่าความผิดปกติของข้อมูลได้ง่ายที่สุด เนื่องจากมีขอบเขตล่างและขอบเขตบนที่ชัดเจน อย่างไรก็ตามตัวสถิติที่ Weiss (1996) ใช้นั้นมีการคำนวณยุ่งยาก ซับซ้อน ต้องมีการอินทิเกรตหลายชั้น ซึ่ง Weiss (1996) แก้ปัญหาโดยการสร้างแบบจำลองโดยวิธีมอนติ คาร์โล (Monte Carlo simulation) หรือจำลองแบบการสุ่มตัวอย่างแบบมีชั้นภูมิ (Stratified sampling simulation) หรือ Gibbs sampling จากการแจกแจงภายหลัง และ Weiss (1996) ได้ให้วิธีการและตัวสถิติไว้กว้าง ๆ ในกรณีของการตรวจสอบค่าผิดปกติบนตัวแบบเชิงเส้น (Linear model)

การวิเคราะห์ความอ่อนไหวแบบเบย์สำหรับการแจกแจงแบบปกติ จำเป็นต้องอินทิเกรต 2 ชั้น (Double integration) เนื่องจากการแจกแจงแบบปกติมี 2 พารามิเตอร์ โดยการอินทิเกรตดังกล่าว จะใช้เทคนิคการวิเคราะห์เชิงตัวเลขในการอินทิเกรต (Numerical integration) ซึ่งมีความยุ่งยากในการคำนวณ อย่างไรก็ตาม หากใช้วิธีการจำลองตาม Weiss (1996) ในการแจกแจงแบบปกติจะพบว่าวิธีการดังกล่าวยังยุ่งยากซับซ้อนเช่นกัน ผู้วิจัยจึงสนใจว่าถ้าลดการอินทิเกรตลงไป 1 ขั้นตอน โดยสมมติว่าทราบความแปรปรวนของประชากร (σ^2) และมอง σ^2 เป็นตัวแปร จากนั้นประมาณความแปรปรวนของประชากร (σ^2) ด้วยค่าประมาณความแปรปรวนที่คำนวณได้จากความแปรปรวนของตัวอย่าง (S^2) แทนการอินทิเกรต 2 ชั้นหรือการจำลองแบบและใช้การอินทิเกรตแค่ชั้นเดียว ตัวสถิติดังกล่าวจะยังใช้ตรวจสอบค่าผิดปกติในชุดข้อมูลที่มาจากประชากรมีที่มีการแจกแจงแบบปกติได้หรือไม่ ผู้วิจัยจึงเสนอวิธีการประมาณตัวสถิติสำหรับตรวจสอบค่าผิดปกติตามแนวทางเบย์ ดังนี้ (1) ประมาณตัวสถิติระยะทางคูลแบคก์ และ (2) ประมาณตัวสถิติระยะทาง L_1

อีกปัญหาหนึ่งสำหรับการวิเคราะห์ความอ่อนไหวแบบเบย์ในการตรวจสอบค่าผิดปกติในข้อมูลทั่วไป คือ การไม่มีความรู้สารสนเทศเกี่ยวกับพารามิเตอร์ก่อนเก็บข้อมูล หากเราไม่ทราบการแจกแจงก่อนของพารามิเตอร์หรือไม่มีความรู้เกี่ยวกับพารามิเตอร์อยู่เลย ผู้วิจัยเห็นว่าสามารถใช้การแจกแจงก่อนแบบไม่มีสารสนเทศตามแนวคิดของ Jeffrey แทนความไม่รู้ดังกล่าว ดังนั้นการวิเคราะห์ความอ่อนไหวแบบเบย์เมื่อใช้การแจกแจงก่อนแบบไม่มีสารสนเทศ การประเมินค่าผิดปกติจะขึ้นอยู่กับข้อมูลเพียงอย่างเดียว ในลักษณะนี้จะคล้ายการตรวจสอบค่าผิดปกติแบบดั้งเดิม ผู้วิจัยจึงอยากเปรียบเทียบกรณีของการตรวจสอบค่าผิดปกติโดยใช้การวิเคราะห์ความอ่อนไหวแบบเบย์กับการตรวจสอบค่าผิดปกติด้วยวิธีดั้งเดิม

วัตถุประสงค์การวิจัย

1. เสนอตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติจากข้อมูลที่มีการแจกแจงแบบปกติ ตัวสถิตินี้มีพื้นฐานมาจากวิธีการแบบเบย์ และได้จากการประมาณตัวสถิติระยะทางคูลแบคก์ และตัวสถิติระยะทาง L_1 ที่ Weiss (1996) ได้เสนอไว้ โดยตัวสถิติที่ได้จากการประมาณตัวสถิติระยะทางคูลแบคก์ คือ ตัวสถิติ $\widetilde{K}$ และตัวสถิติที่ได้จากการประมาณตัวสถิติระยะทาง L_1 คือ ตัวสถิติ $\widetilde{L}_1$ และตัวสถิติ $\acute{L}_1$

2. เปรียบเทียบผลของการตรวจสอบค่าผิดปกติ จากวิธีการวิเคราะห์ความอ่อนไหวแบบเบย์โดยตัวสถิติที่พัฒนาได้จากการวิจัย คือ ตัวสถิติ $\widetilde{K}$, $\widetilde{L}_1$ และ $\acute{L}_1$ กับการใช้วิธีการตรวจสอบค่าผิดปกติแบบหลายค่าซึ่งใช้ตัวสถิติดั้งเดิม คือ ตัวสถิติ Generalized Extreme Studentized Deviate (GESD)

ขอบเขตการวิจัย

1. ข้อมูลที่นำมาศึกษา

1.1 จำลองแบบข้อมูลตัวอย่างจากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน ใช้ขนาดตัวอย่าง 10 , 20 , 50 และ 80 โดยกำหนดขนาดค่าผิดปกติที่ 3 ระดับ คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ นั่นคือ 3σ , 4σ และ 6σ (โดยที่ σ = 1) ตามลำดับ โดยในแต่ละขนาดตัวอย่างจะทำซ้ำ 2,000 รอบ

1.2 ศึกษาและตรวจสอบค่าผิดปกติในชุดข้อมูลจริงโดยมีข้อตกลงเบื้องต้นว่าข้อมูลจริงทั้งหมดที่นำมาศึกษาเป็นข้อมูลมาจากประชากรที่มีการแจกแจงแบบปกติ ชุดข้อมูลจริงมีดังนี้

1.2.1 ข้อมูลของ Freeman (Freeman's Data Sets) (Freeman , quoted in Pettit 1992)

1.2.2 ข้อมูลของDarwin (Darwin's Data) (Fisher 1960)

1.2.3 ข้อมูลของ Sacks et al. (Sacks , Omish , Rosner , Kass and Mclanahan Data) (Sacks et al. , quoted in Rosner 1983)

2. ตัวสถิติที่ใช้ในงานวิจัยประกอบด้วยตัวสถิติทั้งหมด 4 ตัว คือ

2.1 ตัวสถิติ Generalized Extreme Studentized Deviate (GESD)

2.2 ตัวสถิติ $\widetilde{K}$ ที่ได้จากการประมาณตัวสถิติคูลแบคก์ โดยผู้วิจัยได้ประมาณโดยแทนความแปรปรวนของประชากรที่ไม่ทราบค่าด้วยความแปรปรวนของตัวอย่าง

2.3 ตัวสถิติ $\widetilde{L}_1$ และ $\acute{L}_1$ ที่ได้จากการประมาณตัวสถิติระยะทางแบบ L_1 เช่นเดียวกับตัวสถิติคูลแบคก์ ผู้วิจัยได้ประมาณโดยแทนความแปรปรวนของประชากรที่ไม่ทราบค่าด้วยความแปรปรวนของตัวอย่าง ความแตกต่างของตัวสถิติ $\widetilde{L}_1$ และ $\acute{L}_1$ เกิดรูปแบบในการประมาณความแปรปรวนที่แตกต่างกัน

3. ฟังก์ชันความน่าจะเป็นของการแจกแจงก่อนที่ใช้ในการวิจัยครั้งนี้ จะใช้ฟังก์ชันความน่าจะเป็นที่การแจกแจงก่อนเป็นแบบไม่มีสารสนเทศตามแนวคิดของ Jeffrey

4. ฟังก์ชันภาวะน่าจะเป็น เนื่องจากข้อมูลที่นำมาศึกษาเป็นข้อมูลมาจากประชากรที่มีการแจกแจงแบบปกติค่าเฉลี่ยเท่ากับ μ และความแปรปรวนเท่ากับ σ^2 ดังนั้น ฟังก์ชันภาวะน่าจะเป็นที่ใช้ในการวิจัย คือ ฟังก์ชันภาวะน่าจะเป็นของ μ และ σ^2 โดยที่การแจกแจงภายหลังของ μ คือ $N(\bar{x}, \sigma^2)$ และการแจกแจงภายหลังของ σ^2 คือ $\frac{(n-1)S^2}{\chi^2_{n-1}}$

5. ตัวแบบที่ใช้ในการวิจัย เป็นฟังก์ชันความน่าจะเป็นของการแจกแจงภายหลังที่เกิดจากผลคูณของการแจกแจงก่อนแบบไม่มีสารสนเทศ และภาวะน่าจะเป็นจากการแจกแจงแบบปกติของข้อมูล โดยแบ่งตัวแบบออกเป็น 2 ลักษณะ คือ

5.1 ตัวแบบสมบูรณ์ (Complete model) เป็นตัวแบบที่เกิดจากผลคูณของการแจกแจงก่อนแบบไม่มีสารสนเทศ และภาวะน่าจะเป็นจากการแจกแจงแบบปกติของข้อมูลที่ใช้ค่าสังเกตครบทุกค่า

5.2 ตัวแบบที่ถูกก่อกวน (Perturbation model) เป็นตัวแบบที่เกิดจากผลคูณของการแจกแจงก่อนแบบไม่มีสารสนเทศ และภาวะน่าจะเป็นจากการแจกแจงแบบปกติของข้อมูลที่ตัดค่าสังเกตออก 1 ค่า

6. เกณฑ์ที่ใช้ในการพิจารณาค่าผิดปกติ

6.1 สำหรับตัวสถิติที่อิงระเบียบวิธีการแบบดั้งเดิม คือ ตัวสถิติ GESD จะพิจารณาตาม Rosner (1983) ที่นัยสำคัญที่ระดับ 0.05

6.2 สำหรับตัวสถิติที่อิงระเบียบวิธีการแบบเบย์ คือ ตัวสถิติ $\widetilde{K}$, $\widetilde{L}_1$ และ $\acute{L}_1$ โดยส่วนใหญ่จะพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต ถ้าระยะทางของค่าสังเกตใดมีระยะทางห่างจากระยะทางของค่าสังเกตอื่นมาก จะพิจารณาค่าสังเกตนั้นเป็นค่าผิดปกติ แต่เนื่องจากในการจำลองแบบข้อมูล ไม่สามารถที่จะดูระยะทางของตัวสถิติที่เกิดจากค่าสังเกตได้ทั้งหมด ผู้วิจัยจึงใช้อัตราส่วนระยะทางแบบเบย์ในการตัดสินค่าผิดปกติ ในกรณีที่ข้อมูลมีการจำลองแบบ

แนวทางการวิจัย

1. จัดหาข้อมูลที่จะนำมาตรวจสอบค่าผิดปกติ โดย

1.1 จำลองชุดข้อมูลโดยโปรแกรมคอมพิวเตอร์ที่ขนาดตัวอย่าง 4 ขนาด คือ 10 , 20 , 50 และ 80 จากนั้นปะปนค่าผิดปกติลงในชุดข้อมูลดังกล่าว และ

1.2 รวบรวมชุดข้อมูลที่จะนำมาตรวจสอบค่าผิดปกติ ภายใต้ข้อตกลงที่ว่าชุดข้อมูลทุกชุดที่ใช้ตรวจสอบค่าผิดปกติมาจากประชากรที่แจกแจงแบบปกติ

2. ใช้ตัวสถิติ GESD ตรวจหาค่าผิดปกติในชุดข้อมูลทุกชุด ที่ระดับนัยสำคัญ 0.05

3. ใช้ตัวสถิติ $\widetilde{K}$, $\widetilde{L}_1$ และ $\acute{L}_1$ ที่ผู้วิจัยพัฒนาขึ้นตรวจหาค่าผิดปกติในชุดข้อมูลทุกชุด

4. สรุปผลของการตรวจสอบค่าผิดปกติจากผลการตรวจสอบค่าผิดปกติ ในข้อมูลจำลองแบบและข้อมูลจริงที่นำมาศึกษา ของตัวสถิติดั้งเดิม (GESD) และตัวสถิติจากวิธีการวิเคราะห์ความอ่อนไหวแบบเบย์ที่พัฒนาขึ้น คือ ตัวสถิติ $\widetilde{K}$, $\widetilde{L}_1$ และ $\acute{L}_1$

ประโยชน์ของการวิจัย

1. มีทางเลือกใหม่ในการตรวจสอบค่าผิดปกติในข้อมูลปกติทั่วไปหลากหลายขึ้น โดยมีการประยุกต์ตัวสถิติซึ่งพัฒนาจากการวิจัย

2. ลดความยุ่งยากในแง่คำนวณค่าของตัวสถิติที่ได้จากวิธีการวิเคราะห์ความอ่อนไหวแบบเบย์ จากการที่ต้องอินทิเกรต 2 ชั้นเป็นการอินทิเกรตแค่ชั้นเดียว การคำนวณค่าสถิติในวิเคราะห์ความอ่อนไหวแบบเบย์จะทำได้ง่ายขึ้น

3. ทราบความสามารถในการตรวจสอบค่าผิดปกติ ระหว่างตัวสถิติที่อิงระเบียบวิธีแบบดั้งเดิม คือ ตัวสถิติ GESD และตัวสถิติที่อิงระเบียบวิธีแบบเบย์ คือ ตัวสถิติ $\widetilde{K}$, $\widetilde{L}_1$ และ $\acute{L}_1$ ในข้อมูลที่จำลองแบบขึ้น และทำให้ทราบว่าการตรวจสอบค่าผิดปกติในชุดข้อมูลจริงที่นำมาศึกษา จากตัวสถิติ GESD และตัวสถิติ $\widetilde{K}$, $\widetilde{L}_1$ และ $\acute{L}_1$ นำไปสู่ข้อสรุปเช่นเดียวกันหรือแตกต่างกันอย่างไร

บทที่ 2
ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและวรรณกรรมที่เกี่ยวข้องกับการวิจัย ซึ่งมีสาระสามารถสรุปเป็นหมวดหมู่ ได้ดังนี้

1. การแจกแจงที่เกี่ยวข้องกับการวิจัย ประกอบด้วย การแจกแจงแบบปกติ การแจกแจงแบบที การแจกแจงแบบไคสแควร์ และการแจกแจงแบบอินเวอร์สไคสแควร์

2. ทฤษฎี วิธีการและงานวิจัยที่เกี่ยวข้องกับตัวสถิติที่อิงแนวคิดแบบเบย์ ประกอบด้วย กฎของเบย์ การอนุมานทางสถิติแบบเบย์ การเลือกการแจกแจงก่อน การวิเคราะห์ความอ่อนไหวแบบเบย์ ฟังก์ชันก่อกวน ตัวสถิติที่ใช้วัดความอ่อนไหว และการประเมินอิทธิพลของตัวก่อกวน

3. ทฤษฎี วิธีการและงานวิจัยที่เกี่ยวข้องกับตัวสถิติที่อิงแนวคิดแบบดั้งเดิม ประกอบด้วย การตรวจสอบค่าผิดปกติ 1 ค่าและ 2 ค่า การตรวจสอบค่าผิดปกติทีละหลายค่า ตัวสถิติ Generalized Extreme Studentized Deviate (GESD) รูปแบบการพิจารณาค่าผิดปกติของวิธีการ GESD และค่าวิกฤตของตัวสถิติ GESD

การแจกแจงที่เกี่ยวข้องกับการวิจัย

การแจกแจงแบบปกติ (Normal Distribution)

ตัวแปรสุ่ม X ซึ่งมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 จะมีฟังก์ชันความน่าจะเป็น คือ

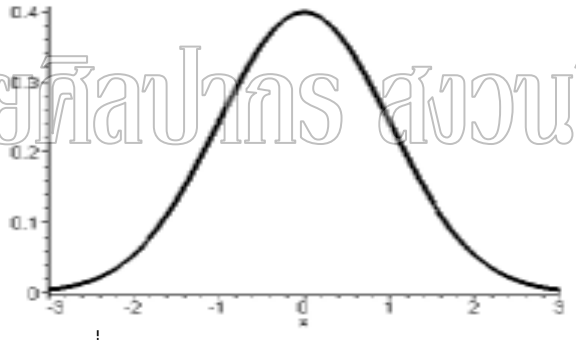
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad , \ -\infty < x < \infty \qquad (2.1)$$

กราฟของฟังชันความหนาแน่นเป็นรูประฆังคว่ำ ดังรูปที่ 1

ถ้าพารามิเตอร์ของการแจกแจงแบบปกติมีค่าของ $\mu = 0$ และ $\sigma^2 = 1$ จะเรียกการแจกแจงแบบปกติมาตรฐาน ซึ่งมีฟังก์ชันความน่าจะเป็นดังนี้

$$f(z) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{z^2}{2}\right] \quad , -\infty < z < \infty \qquad (2.2)$$

การแจกแจงแบบปกติสามารถที่จะเปลี่ยนเป็นรูปมาตรฐานได้ โดยแทน x ใน (2.1) ด้วย $x = \sigma z + \mu$ การแจกแจงแบบปกติถูกค้นพบเมื่อปี ค.ศ. 1733 โดย Abraham de Moivre นักคณิตศาสตร์ชาวฝรั่งเศสแต่ไม่ได้นำมาใช้อย่างแพร่หลายจนกระทั้ง Pierrs S. de Laplace และ Carl F. Gauss ได้พบการแจกแจงแบบปกติโดยแต่ละฝ่ายไม่ทราบผลงานของ Abraham de Moivre มาก่อน ทั้ง Pierrs S. de Laplace และ Carl F. Gauss ได้ทำการศึกษาการแจกแจงของความคลาดเคลื่อนในการวัดทางวิทยาศาสตร์กายภาพด้วยการวัดซ้ำ ๆ กัน และพบว่าผลของการแจกแจงเป็นการแจกแจงแบบปกติ เหตุที่เรียกการแจกแจงแบบนี้ว่าการแจกแจงแบบปกติเพราะในทางปฏิบัติพบว่ามีปรากฏการณ์ทางธรรมชาติหลายปรากฏการณ์มีการแจกแจงแบบปกติ เช่น ความสูงของคน ความเร็วของโมเลกุลของก๊าซในทิศทางใด ๆ ความผิดพลาดจากการวัดต่าง ๆ ในบางครั้งจะเรียกการแจกแจงแบบปกตินี้ว่า การแจกแจงของลาปลาส (Laplacian distribution) หรือการแจกแจงของเกาส์ (Gaussian distribution) เพื่อเป็นเกียรติแก่ Pierrs S. de Laplace และ Carl F. Gauss (Stigler 1986 ; Weiss 1993)



รูปที่ 1 กราฟการแจกแจงแบบปกติ

พิจารณากราฟของการแจกแจงแบบปกติ (รูปที่ 1) โค้งความถี่ของข้อมูลที่มีการแจกแจงแบบปกติจะต้องมีลักษณะสมมาตรและต้องไม่สูงมากไป ไม่ต่ำจนเกินไป การตรวจสอบข้อมูลว่ามีลักษณะแจกแจงแบบปกติหรือไม่ ก็จะพิจารณาจากลักษณะทั้งสองอย่างดังกล่าว ข้อมูลที่แจกแจงแบบปกติจะมีลักษณะสมมาตรรอบค่าเฉลี่ย ในข้อมูลที่มีการแจกแจงแบบปกติสามารถวัดความเบ้ของโค้งความถี่ของข้อมูลได้หลายวิธี เช่น วิธีของเพียร์สัน วิธีของบาวเลย์ วิธีโมเมนต์ เป็นต้น (Box and Tiao 1973) นอกเหนือจากลักษณะสมมาตรแล้วการกระจายของข้อมูลต้องไม่มากและไม่น้อยเกินไป หรือโค้งความถี่จะไม่สูงเกินไปและไม่เตี้ยเกินไป ซึ่งเรียกว่า โด่งปานกลาง (Mesokurtic) ส่วนโค้งความถี่ที่มีลักษณะสมมาตรแต่ความโด่งไม่อยู่ในระดับปานกลางก็จะไม่ใช่การแจกแจงแบบปกติ โค้งความถี่ลักษณะสมมาตรที่ความโด่งของความ

โค้งอยู่ในระดับสูงมาก จะเรียกว่า โด่งมาก (Leptokurtic) ถ้าความโด่งอยู่ในระดับเตี้ย จะเรียกว่า โด่งน้อย (Platykurtic) นอกจากนี้ในชีวิตประจำวันมักจะพบการแจกแจงแบบปกตินี้เสมอ ๆ ทั้งนี้เนื่องมาจากทฤษฎีลิมิตเข้าสู่ส่วนกลาง (Central limit theorem) ซึ่งกล่าวว่าภายใต้เงื่อนไขที่เหมาะสม ถ้านำตัวแปรสุ่ม ซึ่งไม่จำเป็นต้องมีการแจกแจงแบบปกติจำนวนมากมาบวกกันแล้ว ผลบวกของตัวแปรดังกล่าวโดยประมาณจะมีการแจกแจงแบบปกติ กล่าวคือ ถ้า $X_1,X_2,...,X_n$ เป็นอิสระกันและมีการแจกแจงแบบเดียวกัน (Independent and identically distribution ซึ่งเรียกย่อๆว่า iid) โดยมี $E(X) = \mu$ และ $Var(X) = \sigma^2 < \infty$ แล้ว

$$\lim_{n\to\infty} \Pr\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)\le x\right] = \Phi(x)$$

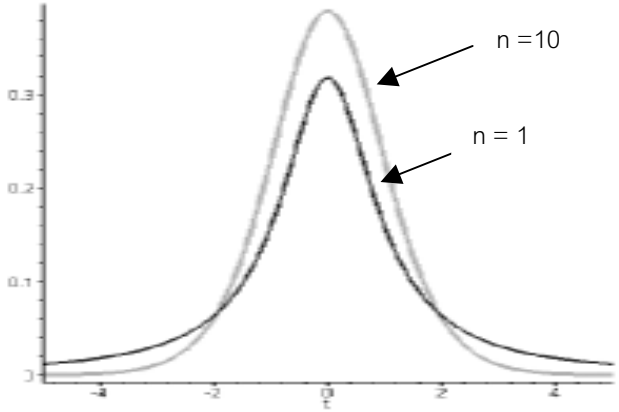
เมื่อ $\Phi(x)$ เป็นฟังก์ชันการแจกแจงของตัวแปรสุ่มที่มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1 สัญลักษณ์ที่ใช้แทนการแจกแจงแบบปกติที่ค่าเฉลี่ย μ และความแปรปรวน σ^2 คือ $N(\mu,\sigma^2)$

การแจกแจงแบบที (t – Distribution)

ตัวแปรสุ่ม T เป็นตัวแปรสุ่มที่ต่อเนื่องและมีฟังก์ชันความน่าจะเป็น คือ

$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty$$

เรียก T ว่ามีการแจกแจงแบบที ที่มีองศาอิสระ (Degree of freedom) เท่ากับ n กราฟของฟังก์ชันความน่าจะเป็นของการแจกแจงแบบที ที่มีองศาอิสระเท่ากับ 1 และ 10 ดังรูปที่ 2



รูปที่ 2 กราฟการแจกแจงแบบที ที่มีองศาอิสระเท่ากับ 1 และ 10

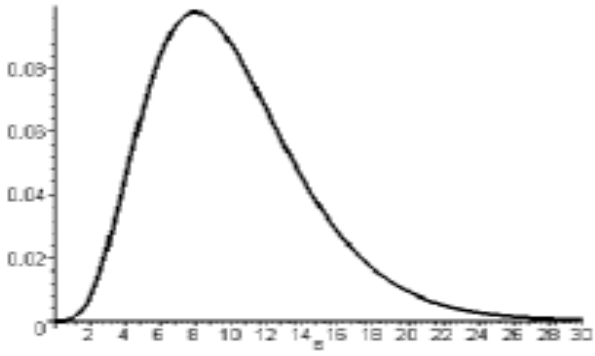
การแจกแจงแบบทีค้นพบโดย William Sealy Gosset ในปี ค.ศ. 1908 ข้อเขียนเกี่ยวกับการแจกแจงแบบทีได้ลงพิมพ์เป็นบทความเรื่อง “The Probable Error of a Mean” ในวารสาร Biometrika โดยใช้นามปากกาว่า สติวเดนต์ (Student) ดังนั้นการแจกแจงแบบทีจึงมีอีกชื่อหนึ่งว่า การแจกแจงแบบสติวเดนต์ – ที (Student’s t Distribution) (Fisher 1990 ; Weiss 1993 ; O'Connor and Robertson 2002) การแจกแจงแบบทีจะมีลักษณะสมมาตรและเป็นรูประฆังคว่ำ (รูปที่ 2) มีจุดสูงสุดของโค้งเพียงจุดเดียว (Unimodel) สัญลักษณ์ที่ใช้แทนการแจกแจงแบบทีที่มีองศาอิสระเท่ากับ n คือ t_n โดยมีค่าเฉลี่ยและความแปรปรวนเท่ากับ 0 และ $\frac{n}{n-2}$ ตามลำดับ ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยมในการแจกแจงแบบทีจะมีค่าเท่ากัน ถ้าตัวอย่างมีขนาดเพิ่มขึ้น การแจกแจงแบบทีจะมีการแจกแจงที่เข้าใกล้การแจกแจงแบบปกติ นั่นคือ เมื่อ $n \to \infty$ การแจกแจงแบบทีจะเข้าใกล้การแจกแจงแบบปกติมาตรฐาน

การแจกแจงแบบไคสแควร์ (Chi – square Distribution)

ตัวแปรสุ่ม Y ที่เป็นตัวแปรสุ่มที่ต่อเนื่องและมีฟังก์ชันความน่าจะเป็น คือ

$$f(y) = \frac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)}\, y^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}y\right), \quad y > 0$$

เรียก Y ว่ามีการแจกแจงแบบไคสแควร์ที่มีองศาอิสระเท่ากับ n กราฟของฟังก์ชันความน่าจะเป็นของการแจกแจงแบบไคสแควร์ ที่มีองศาอิสระเท่ากับ 10 ดังรูปที่ 3



รูปที่ 3 กราฟการแจกแจงแบบไคสแควร์ ที่มีองศาอิสระเท่ากับ 10

ในปี ค.ศ. 1875 Helmert ได้ค้นพบการแจกแจงแบบไคสแควร์และในปี ค.ศ. 1900 Karl Pearson ได้คิดค้นการแจกแจงแบบไคสแควร์เพื่อทดสอบภาวะสารูปดี (Test for

goodness of fit) และได้เผยแพร่การแจกแจงแบบไคสแควร์เป็นครั้งแรก (Stigler 1986 ; Fisher 1990 ; Weiss 1993) สัญลักษณ์ χ อ่านว่า ไค (Chi) เป็นอักษรกรีก ดังนั้น χ^2 อ่านว่า ไคสแควร์ (Chi – square) จึงเป็นสัญลักษณ์ที่แทนการแจกแจงแบบไคสแควร์ โดยที่ χ^2_n หมายถึง การแจกแจงแบบไคสแควร์ที่มีองศาอิสระเท่ากับ n การแจกแจงแบบไคสแควร์จะมีลักษณะที่ไม่สมมาตร (Non symmetry) โดยมีจุดสูงสุดเพียงจุดเดียว ค่าเฉลี่ย ความแปรปรวน ค่ามัธยฐาน และค่าความเบ้ของการแจกแจงแบบไคสแควร์ที่มีองศาอิสระ n เท่ากับ n 2n n-2 และ $\sqrt{\frac{8}{n}}$ ตามลำดับ นอกจากนี้การแจกแจงแบบไคสแควร์ยังมีความสัมพันธ์กับการแจกแจงแบบปกติ กล่าวคือ ถ้า X มีการแจกแจงแบบปกติมาตรฐานและ $Y = X^2$ แล้ว Y จะมีการแจกแจงแบบไคสแควร์ที่มีองศาอิสระเท่ากับ 1

การแจกแจงแบบอินเวอร์สไคสแควร์ (Inverted Chi – square Distribution)

ตัวแปรสุ่ม Q เป็นตัวแปรสุ่มที่ต่อเนื่องและมีฟังก์ชันความน่าจะเป็น คือ

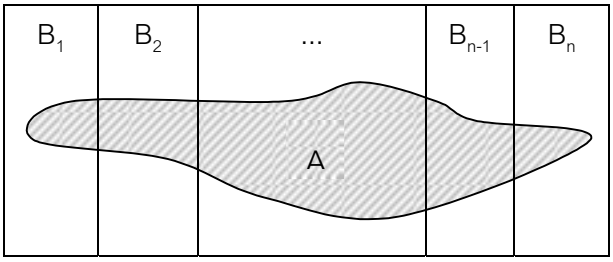
$$f(q) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}(q)^{-[\frac{n}{2}-1]}\exp(-\frac{1}{2q}) \quad , \quad q > 0$$

เรียก q ว่ามีการแจกแจงแบบอินเวอร์สไคสแควร์ ที่มีองศาอิสระเท่ากับ n โดยส่วนกลับของ Q จะมีการแจกแจงแบบไคสแควร์ (Box and Tiao 1973) การแจกแจงแบบอินเวอร์สไคสแควร์มีลักษณะที่ไม่สมมาตร โดยการแจกแจงแบบอินเวอร์สไคสแควร์ ที่มีองศาอิสระเท่ากับ n จะมีค่าเฉลี่ยและความแปรปรวนเท่ากับ $\frac{1}{n}$ และ $\frac{1}{2n}$ ตามลำดับ การวิเคราะห์แบบเบย์ (Bayesian analysis) สำหรับการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 การแจกแจงแบบอินเวอร์สไคสแควร์นับว่ามีความสำคัญมาก เพราะ σ^2 มีการแจกแจงแบบอินเวอร์สไคสแควร์

ทฤษฎี วิธีการและงานวิจัยที่เกี่ยวข้องกับตัวสถิติที่อิงแนวคิดแบบเบย์

กฎของเบย์ (Bayes' Rule)

ถ้าทำการแบ่งส่วนสเปซตัวอย่าง S ออกเป็นส่วนๆ (Partition) สมมติเป็น n ส่วน เรียก $B_1,B_2,...,B_n$ แต่ละส่วนไม่มีการซ้ำซ้อนกัน $B_i \cap B_j = \phi$ เมื่อ $i \neq j$ และนอกจากนี้ B_i ทั้งหมดจะประกอบกันเป็น S พอดี นั่นคือ $B_1 \cup B_2 \cup ... \cup B_n = S$ และไม่มี B_i ตัวใดเป็นเซตว่าง $B_i \neq \phi$ ทุกค่าของ i , i = 1 , 2 , ... , n ถ้าทราบค่าของ $\Pr(B_i)$ ทั้งหมด เมื่อ A เป็นเหตุการณ์ที่สนใจและเกิดขึ้นแล้ว กฎของเบย์จะช่วยให้หาค่าของ $\Pr(B_i|A)$ ได้



รูปที่ 4 กฎของเบย์

จากภาพในรูปที่ 4

$S = B_1 \cup B_2 \cup ... \cup B_n$

$\Pr(B_i) \neq 0$, $\Pr(B_i \cap B_j) = 0$, $i \neq j$

$A = (A \cap B_1) \cup (A \cap B_2) \cup ... \cup (A \cap B_n)$

เนื่องจาก $B_i \cap B_j = \phi$ เมื่อ $i \neq j$ ดังนั้น $A \cap B_1 , A \cap B_2 , ... , A \cap B_n$ จึงเป็นเซตที่ไม่มีสมาชิกร่วมกัน

$$\begin{aligned}\Pr(A) &= \Pr(A \cap B_1) + \Pr(A \cap B_2) + ... + \Pr(A \cap B_n) \\ &= \Pr(A|B_1)\Pr(B_1) + \Pr(A|B_2)\Pr(B_2) + ... + \Pr(A|B_n)\Pr(B_n) \\ &= \sum_{i=1}^{n} \Pr(A|B_i)\Pr(B_i)\end{aligned}$$

ถ้ากำหนดว่าเหตุการณ์ A ได้เกิดขึ้นแล้ว โอกาสที่จะเกิดเหตุการณ์ B_i จะเป็น

$$\Pr(B_j|A) = \frac{\Pr(A \cap B_j)}{\Pr(A)} = \frac{\Pr(A|B_j)\Pr(B_j)}{\sum_{i=1}^{n} \Pr(A|B_i)\Pr(B_i)} \qquad (2.3)$$

ถ้าพิจารณาว่า B_i เป็นสถานะของธรรมชาติ (States of nature) ที่เป็นไปได้ ในแง่นี้ $\Pr(A|B_i)$ อาจถือว่าเป็นโอกาสที่จะเกิด A เมื่อสถานะของธรรมชาติ คือ B_i นั่นก็คือ ความน่าจะเป็นแบบมีเงื่อนไขของ A เมื่อกำหนด B_i ส่วน $\Pr(B_i)$ ก็คือ ความน่าจะเป็นของสถานะธรรมชาติแต่ละ

สถานะก่อนทำการทดลอง และเรียกกันว่า ความน่าจะเป็นก่อน (Prior probability) และ $\Pr(B_i|A)$ อาจเรียกว่าเป็นความน่าจะเป็นใหม่เกี่ยวกับธรรมชาติหลังจากได้ทำการทดลองไปแล้ว และพบว่าเหตุการณ์ A เกิดขึ้นจึงเรียก $\Pr(B_i|A)$ว่าความน่าจะเป็นภายหลัง (Posterior probability)

การอนุมานทางสถิติแบบเบย์ (Statistical Inference of Bayesian)

Thomas Bayes นักปราชญ์ชาวอังกฤษผู้เสนอทฤษฎีเบย์ (Bayes' theorem) ซึ่งมีประโยชน์ในการอธิบายการทดลองที่เป็น 2 ขั้นตอน โดยที่ผลลัพธ์ที่ได้จากการทดลองขั้นที่ 2 ขึ้นอยู่กับผลจากการทดลองขั้นแรกด้วย การทดลองแบบนี้เรียกว่า การทดลองเชิงประกอบ (Compound experiment) แนวคิดการอนุมานทางสถิติโดยอาศัยทฤษฎีเบย์เป็นแนวคิดใหม่และพัฒนานำไปประยุกต์ใช้ โดยเฉพาะอย่างยิ่งทางปัญหาการตัดสินใจ (Decision problem) แนวความคิดการอนุมานนี้ถือว่า พารามิเตอร์เป็นตัวแปรสุ่มตัวหนึ่ง ดังนั้น จึงมีตัวแบบความน่าจะเป็นที่สามารถอธิบายความเชื่อเกี่ยวกับพารามิเตอร์ดังกล่าวได้ แนวความคิดในการอนุมานทางสถิติแบบเบย์มีดังนี้

ให้ θ เป็นพารามิเตอร์ที่สนใจซึ่งเป็นสมาชิกใน Θ

$p(\theta)$ เป็นฟังก์ชันความน่าจะเป็นบน Θ ซึ่งแทนความเชื่อที่มีอยู่เกี่ยวกับคุณลักษณะของ θ ก่อนที่จะทำการเก็บตัวอย่างมาสังเกตโดยมีคุณสมบัติว่า สำหรับจำนวน a ,b ใด ๆ

- $\Pr(a \le \theta \le b) = \int_a^b p(\theta)\,d\theta$ เมื่อ θ เป็นตัวแปรสุ่มชนิดต่อเนื่อง
- $\Pr(a \le \theta \le b) = \sum_{\theta=a}^{b} p(\theta)$ เมื่อ θ เป็นตัวแปรสุ่มชนิดไม่ต่อเนื่อง

จะเรียก $p(\theta)$ นี้ว่า การแจกแจงก่อน (Prior distribution) ของ θ หลังจากเก็บตัวอย่างสุ่ม $\underline{X} = (X_1, X_2, ..., X_n)$ ขนาด n ซึ่งมีฟังก์ชันความน่าจะเป็นร่วมเป็น $p(\underline{x};\theta)$ มา สำหรับแนวคิดของเบย์ $p(\underline{x};\theta)$ นี้จะเป็นฟังก์ชันความน่าจะเป็นแบบมีเงื่อนไขของ $\underline{X}$ เมื่อกำหนด θ ดังนั้น จึงใช้สัญลักษณ์ $p(\underline{x}|\theta)$ แทน

ทฤษฎีที่ 1 ทฤษฎีของเบย์ (Bayes' theorem) (Box and Tiao 1973)

ให้ $\underline{X}$ = (X_1 , X_2 , ... , X_n) เป็นตัวอย่างสุ่มขนาด n และมีฟังก์ชันความน่าจะเป็น $p(\underline{x}|\underline{\theta})$ ขึ้นกับพารามิเตอร์ $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ k ค่า สมมติว่า $\underline{\theta}$ มีฟังก์ชันการแจกแจงความน่าจะเป็น คือ $p(\underline{\theta})$ เมื่อ $p(\underline{x}|\underline{\theta})\ p(\underline{\theta}) = p(\underline{x}, \underline{\theta}) = p(\underline{\theta}|\ \underline{x})\ p(\underline{x})$ (2.4)

สำหรับค่าสังเกต $\underline{x}$ = (x_1 , x_2 , ... , x_n) การแจกแจงแบบมีเงื่อนไขของ $\underline{\theta}$ เมื่อสังเกต $\underline{x}$ คือ

$$p(\underline{\theta}|\ \underline{x}) = \frac{p(\underline{x}|\ \underline{\theta})p(\underline{\theta})}{p(\underline{x})} \qquad (2.5)$$

โดยที่ $p(\underline{x}) = \int_{\underline{\theta}} p(\underline{x}|\underline{\theta})p(\underline{\theta})d\underline{\theta}$ กรณีที่ θ ต่อเนื่อง

$= \sum_{\underline{\theta}} p(\underline{x}|\underline{\theta})p(\underline{\theta})$ กรณีที่ θ ไม่ต่อเนื่อง

$p(\underline{\theta}|\ \underline{x})$ จะเป็นฟังก์ชันความน่าจะเป็นของ $\underline{\theta}$ ที่อธิบายความเชื่อเกี่ยวกับ $\underline{\theta}$ ภายหลังจากทำการสังเกตข้อมูล $\underline{x}$ = (x_1 , x_2 , ... , x_n) และเรียก $p(\underline{\theta}|\ \underline{x})$ ว่าการแจกแจงภายหลัง (Posterior distribution) ของ $\underline{\theta}$ เมื่อกำหนด $\underline{x}$ กล่าวโดยสรุป แนวคิดของการอนุมานแบบเบย์นี้จะเป็นการใช้ข้อมูลหรือสารสนเทศ ซึ่งมีในตัวอย่างสุ่ม $\underline{X}$ ทั้ง n ค่าสังเกตมาเปลี่ยนแปลงความเชื่อดั้งเดิมเกี่ยวกับ $\underline{\theta}$ โดยการเปลี่ยนแปลงการแจกแจงก่อนของ $\underline{\theta}$ ไปสู่การแจกแจงภายหลังของ $\underline{\theta}$

เมื่อมีข้อมูล x_1 , x_2 , ... , x_n และมี $p(\underline{x}|\underline{\theta})$ อาจถือว่า $p(\underline{x}|\underline{\theta})$ เป็นฟังก์ชันของพารามิเตอร์ $\underline{\theta}$ ตามแนวคิดของ Fisher (Box and Tiao 1973) และเรียกฟังก์ชันดังกล่าวว่าฟังก์ชันภาวะน่าจะเป็นของ $\underline{\theta}$ เมื่อทราบ x_1 , x_2 , ... , x_n ซึ่งแทนด้วย $L(\underline{\theta}|\ \underline{x})$ จึงสรุปว่า

$$p(\underline{\theta}|\ \underline{x}) = \frac{L(\underline{\theta}|\ \underline{x})p(\underline{\theta})}{p(\underline{x})}$$

นั่นคือ การแจกแจงภายหลัง $\propto$ (ฟังก์ชันภาวะน่าจะเป็น)$\times$ (การแจกแจงก่อน)

{ Posterior Distribution $\propto$ (Likelihood Function) $\times$ (Prior Distribution) }

ฟังก์ชันภาวะน่าจะเป็นของตัวอย่างสุ่ม X_1 , X_2 , ... , X_n ที่มาจากการแจกแจงแบบปกติ กล่าวคือ X_i , i = 1,2,...,n มีการแจกแจง $N(\mu, \sigma^2)$ โดยที่พารามิเตอร์ μ และ σ^2 ไม่ทราบค่า และ X_i แต่ละตัวเป็นอิสระต่อกัน คือ

$$L(\mu, \sigma^2|\ \underline{x}) = L(\mu, \sigma^2|x_1)\ L(\mu, \sigma^2|x_2) \ldots\ L(\mu, \sigma^2|x_n) \qquad (2.6)$$

เมื่อ $$L\left(\mu,\sigma^2\middle|x_i\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-x_i}{\sigma}\right)^2\right] \quad , -\infty<\mu<\infty, \quad \sigma^2>0 \quad (2.7)$$

จะได้ $$L\left(\mu,\sigma^2\middle|\underline{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum\left(\mu-x_i\right)^2\right] \quad , -\infty<\mu<\infty \quad , \sigma^2>0 \quad (2.8)$$

พิจารณา

$$\begin{aligned}\sum(\mu-x_i)^2 &= \sum\left(x_i-\mu\right)^2\\ &= \sum\left(x_i-\bar{x}+\bar{x}-\mu\right)^2\\ &= \sum\left(x_i-\bar{x}\right)^2+\sum\left(\bar{x}-\mu\right)^2\\ &= (n-1)S^2+n\left(\bar{x}-\mu\right)^2\end{aligned}$$

เมื่อ $\bar{x} = \frac{\sum x_i}{n}$ และ $S^2 = \frac{\sum\left(x_i-\bar{x}\right)^2}{n-1}$

ดังนั้น $$L\left(\mu,\sigma^2\middle|\underline{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left[-\frac{1}{2\sigma^2}\left[(n-1)S^2+n\left(\bar{x}-\mu\right)^2\right]\right] \quad (2.9)$$

การเลือกการแจกแจงก่อน (Prior Distribution Selection)

ในการอนุมานทางสถิติแบบเบย์จะเห็นว่าการแจกแจงก่อนของ θ มีบทบาทที่สำคัญ การเลือกรูปแบบการแจกแจงก่อนของ θ ขึ้นอยู่กับว่ามีความรู้เกี่ยวกับ θ อยู่ก่อนที่จะเก็บข้อมูลหรือไม่ ถ้ามีความรู้เกี่ยวกับ θ อยู่ นิยมแทนความรู้เกี่ยวกับ θ ในรูปของการแจกแจงที่เมื่อนำไปรวมกับฟังก์ชันภาวะน่าจะเป็นที่ได้จากข้อมูลที่รวบรวมมาแล้ว สามารถคำนวณได้ง่าย การแจกแจงก่อนที่นิยมใช้ในกรณีนี้เช่น Conjugate prior แต่ถ้าไม่มีความรู้เกี่ยวกับ θ อยู่เลย ได้มีนักสถิติหลายท่านเสนอการแจกแจงที่แทนการไม่มีความรู้เกี่ยวกับ θ ไว้เช่น Jeffreys prior , Laplace prior และ Zellner prior เป็นต้น

1 . Conjugate prior

พิจารณาตัวอย่างสุ่ม X_1 , X_2 , ... , X_n ที่มีฟังก์ชันความน่าจะเป็น $f(x;\theta)$ ฟังก์ชันความน่าจะเป็น Conjugate prior สำหรับพารามิเตอร์ θ ใด ๆ ก็คือ การแจกแจงก่อน $p(\theta)$ ที่มีรูปแบบเช่นเดียวกับการแจกแจงภายหลัง $p(\theta|x)$ แต่มีพารามิเตอร์ต่างไป เช่น ตัวอย่างสุ่มชุดหนึ่งสุ่มจากประชากรที่มีฟังก์ชันความน่าจะเป็นแบบทวินาม ซึ่งมีพารามิเตอร์ β ถ้าการแจกแจงก่อนของ β คือ ฟังก์ชันความน่าจะเป็นแบบเบต้าแล้วการแจกแจงภายหลังของ β ก็จะมีรูปแบบการแจกแจงแบบเบต้าด้วย

Press (1989) สรุปบาง Conjugate prior ที่ใช้กับตัวอย่างสุ่มที่มีฟังก์ชันความน่าจะเป็นต่าง ๆ ดังตารางที่ 1

ตารางที่ 1 การแจกแจง Conjugate prior

ลักษณะฟังก์ชันความน่าจะเป็นของตัวอย่างสุ่ม	การแจกแจง Conjugate Prior
1. Binomial (β)	β ~ Beta
2. Negative Binomial (β)	β ~ Beta
3. Poisson (m)	m ~ Gamma
4. Exponential (λ^{-1})	λ ~ Gamma
5. Normal with known variance but unknown mean (μ)	μ ~ Normal
6. Normal with known mean but unknown variance (σ^2)	σ^2 ~ Inverted Gamma หรือ Inverted Chi - square

2 . Prior Representing Ignorance

บ่อยครั้งที่ไม่แน่ใจในข้อมูลหรือสารสนเทศเกี่ยวกับ θ ก่อนที่จะทำการศึกษาจากข้อมูล $\underline{X}$ หรือบางครั้งต้องการที่จะอนุมานแบบเบย์โดยกระทำในลักษณะที่ไม่ใช้ความรู้เกี่ยวกับ θ ที่มีอยู่เดิม ฟังก์ชันการแจกแจงก่อนของ θ ที่แทนการไม่มีความรู้เกี่ยวกับ θ เรียก การแจกแจงก่อนแบบไม่มีสารสนเทศ (Noninformative prior distribution) ตัวอย่างของการแจกแจงก่อนลักษณะนี้ เช่น

2.1 Jeffreys prior ให้ θ เป็นพารามิเตอร์ในฟังก์ชันความน่าจะเป็น $f(x;\theta)$ ของตัวแปรสุ่ม X ฟังก์ชันความน่าจะเป็นก่อนหรือการแจกแจงก่อนแบบ Jeffreys สำหรับ θ คือ

$$p(\theta) \propto \frac{1}{\sqrt{I(\theta)}} \quad \text{หรือ} \quad p(\theta) = \frac{c}{\sqrt{I(\theta)}}$$

$$\text{เมื่อ} \quad I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right] \quad \text{หรือ} \quad I(\theta) = E\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2$$

และ c เป็นค่าคงที่ใด ๆ

2.2 Laplace prior คือการกำหนดให้ฟังก์ชันความน่าจะเป็นก่อน สำหรับ θ และ $p(\theta)$ มีรูปแบบดังนี้

$$p(\theta) = c \quad , \quad -\infty < \theta < \infty \quad \text{หรือ} \quad p(\theta) = \frac{1}{b-a} \quad , \quad a < \theta < b$$

เมื่อใช้ Laplace prior ฟังก์ชันความน่าจะเป็นภายหลังหรือการแจกแจงภายหลังของ θ จะเป็นสัดส่วนกับฟังก์ชันภาวะน่าจะเป็น เพราะค่าคงที่ c ไม่สำคัญ ดังนั้น ส่วนใหญ่มักกำหนดให้ $p(\theta) = 1 \quad , \quad -\infty < \theta < \infty$

การแจกแจงภายหลัง ภายใต้ภาวะน่าจะเป็นแบบปกติและการแจกแจงก่อนแบบไม่มีสารสนเทศของ Jeffrey

กรณีที่ 1 การแจกแจงแบบปกติ $N(\mu, \sigma^2)$ เมื่อทราบ σ^2 จะได้ฟังก์ชันภาวะน่าจะเป็น คือ

$$L\left(\mu \middle| \sigma^2, \underline{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right], -\infty < \mu < \infty$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right], -\infty < \mu < \infty \qquad (2.10)$$

การแจกแจงก่อนแบบไม่มีสารสนเทศที่ใช้ คือ $p(\mu|\sigma) = C$ เมื่อ C เป็นค่าคงที่ใด ๆ นั่นคือ $\sqrt{I(\theta)} = C$ ดังนั้น จะได้การแจกแจงภายหลัง คือ

$$p\left(\mu \middle| \sigma^2, \underline{x}\right) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right] \quad , \quad -\infty < \mu < \infty \qquad (2.11)$$

จะเห็นว่า $\mu \middle| \sigma^2, \underline{x} \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$

กรณีที่ 2 การแจกแจงแบบปกติ $N(\mu, \sigma^2)$ เมื่อทราบ μ การแจกแจงก่อนแบบไม่มีสารสนเทศในกรณีนี้ คือ $p(\sigma|\mu) = \sigma^{-1}$ และฟังก์ชันภาวะน่าจะเป็น คือ

$$L\left(\sigma^2 \middle| \mu, \underline{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{ns^2}{2\sigma^2}\right] \quad , \quad \sigma > 0 \quad \text{และ} \quad s^2 = \frac{\sum(x_i-\mu)^2}{n}$$

ดังนั้น จะได้การแจกแจงภายหลัง คือ

$$p(\sigma|\mu,\underline{x}) = \frac{(ns^2)^{\frac{n}{2}}\sigma^{-(n+1)}}{2^{\left(\frac{n}{2}\right)-1}\Gamma\left(\frac{n}{2}\right)}\exp\left[-\frac{ns^2}{2\sigma^2}\right] \quad , \quad \sigma>0 \qquad (2.12)$$

จะได้ $\sigma^2|\mu,\underline{x} \sim \frac{(n-1)s^2}{\chi^2_{n-1}}$

กรณีที่ 3 การแจกแจงแบบปกติ $N(\mu,\sigma^2)$ เมื่อไม่ทราบ μ และ σ^2 สำหรับการแจกแจงก่อนแบบไม่มีสารสนเทศที่ใช้ คือ $p(\mu,\sigma) = \sigma^{-1}$ จะได้การแจกแจงภายหลังในกรณีนี้ คือ

$$p(\mu,\sigma|\underline{x}) = \sqrt{\frac{n}{2\pi}}\left[\frac{1}{2}\Gamma\left(\frac{n-1}{2}\right)\right]^{-1}\left(\frac{(n-1)s^2}{2}\right)^{\frac{n-1}{2}}\sigma^{-(n+1)}e^{-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\mu-\bar{x})^2\right]} \quad (2.13)$$

เมื่อ $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$

เนื่องจาก

a) $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

b) $(n-1)s^2 = \sum(x_i-\bar{x})^2 \sim \sigma^2\chi^2_{n-1}$ และ

c) $\bar{x}$ และ s^2 เป็นตัวสถิติที่เป็นอิสระกัน

ดังนั้น $p(\bar{x},s^2|\mu,\sigma^2) = p(\bar{x}|\mu,\sigma^2)p(s^2|\sigma^2)$

โดยที่ $p(\bar{x}|\mu,\sigma^2) = \sqrt{\frac{n}{2\pi}}\sigma^{-1}\exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]$, $-\infty<\bar{x}<\infty$

แ ล

$$p(s^2|\sigma^2) = \frac{1}{\Gamma\left(\frac{1}{2}(n-1)\right)}\left(\frac{n-1}{2\sigma^2}\right)^{\frac{1}{2}(n-1)}(s^2)^{\frac{1}{2}(n-1)-1}\exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right] \quad , \quad s^2>0$$

ส่วนประกอบของการแจกแจง $p(\mu,\sigma^2 \mid \underline{x})$ ซึ่งคือการแจกแจงร่วมภายหลัง (Joint posterior distribution) ของ (μ,σ^2) สามารถเขียนได้เป็น $p(\mu,\sigma^2|\underline{x}) = p(\mu|\sigma^2,\underline{x})\ p(\sigma^2|\underline{x})$ เมื่อส่วนประกอบตัวแรก คือ การแจกแจงภายหลังแบบมีเงื่อนไข (Conditionnal posterior distribution) ของ μ เมื่อกำหนด σ^2 และส่วนประกอบที่สอง คือ การแจกแจงมาร์จินัลภายหลัง (Marginal posterior distribution) ของ σ^2

การวิเคราะห์ความอ่อนไหวแบบเบย์ (Bayesian Sensitivity Analysis)

ดังที่กล่าวมาแล้วว่าในการอนุมานแบบเบย์อนุมานจากการแจกแจงภายหลัง เนื่องจากการแจกแจงภายหลังเกิดจากการแจกแจงก่อนและภาวะน่าจะเป็น ดังนั้น ถ้ามีการเปลี่ยนแปลงในการแจกแจงก่อน หรือในภาวะน่าจะเป็นย่อมจะก่อให้เกิดการเปลี่ยนแปลงในการแจกแจงภายหลัง การวิเคราะห์ความอ่อนไหวแบบเบย์ เป็นการประเมินความอ่อนไหวของการแจกแจงภายหลังที่เกิดจากการเปลี่ยนแปลงในการแจกแจงก่อนหรือในภาวะน่าจะเป็น ถ้าการแจกแจงภายหลังไม่อ่อนไหวต่อการเปลี่ยนแปลงของการแจกแจงก่อน หรือภาวะน่าจะเป็นจะทำให้ได้ผลการอนุมานที่ได้เชื่อถือได้

ฟังก์ชันก่อกวน (Perturbation Function)

ในการวิเคราะห์ความอ่อนไหวแบบเบย์ นิยมประเมินความอ่อนไหวระหว่างตัวแบบสมบูรณ์ (Complete model : M_0) กับตัวแบบที่ถูกก่อกวน (Perturbation model : M_1) ตัวแบบที่ถูกก่อกวนนั้นจะถูกก่อกวนด้วยฟังก์ชันการก่อกวน (Perturbation function : h^*) Kass , Tierney and Kadane (1989) ได้เสนอฟังก์ชันการก่อกวนไว้ 2 ลักษณะ คือ

1. ฟังก์ชันก่อกวนการแจกแจงก่อน (Prior perturbation function) เป็นฟังก์ชันก่อกวนที่ใช้ประเมินอิทธิพลการก่อกวนของการแจกแจงก่อนที่ส่งผลต่อการแจกแจงภายหลัง ถ้าให้ฟังก์ชันการก่อกวน คือ h^* และฟังก์ชันก่อกวนการแจกแจงก่อน คือ h^*_1 โดยนิยามตาม Kass , Tierney and Kadane (1989) จะได้ $h^*_1(\theta) = \{ q(\theta) / p(\theta) \}$ เมื่อ $q(\theta)$ คือ ฟังก์ชันการแจกแจงก่อนใหม่ที่นำไปแทนที่ $p(\theta)$ ซึ่งเป็นฟังก์ชันการแจกแจงก่อนเดิม เมื่อนำฟังก์ชันก่อกวนการแจกแจงก่อนไปก่อกวนตัวแบบสมบูรณ์ จะได้ตัวแบบที่ถูกก่อกวนซึ่งคือการแจกแจงภายหลังที่ได้จากการการแจกแจงก่อนที่ถูกแทนที่ด้วยการแจกแจงก่อนใหม่หรือจะเรียกว่า การแจกแจงภายหลังการก่อกวน ตัวแบบสมบูรณ์สามารถแสดงได้ดังนี้ คือ

$$M_0 : \ p(\theta|\underline{x}) = \frac{L(\theta|\underline{x})p(\theta)}{p(\underline{x})} \qquad (2.14)$$

ก่อกวนด้วยฟังก์ชันก่อกวน $h^{*}_{1}(\theta)$ จะได้

ตัวแบบที่ถูกก่อกวน คือ $M_1 : p_1(\theta|\underline{x}) = \frac{L(\theta|\underline{x})p(\theta)h^{*}_{1}(\theta)}{p_1(\underline{x})} = \frac{L(\theta|\underline{x})q(\theta)}{p_1(\underline{x})}$ (2.15)

เมื่อ $p_1(\underline{x}) = \int_{\theta} L(\theta|\underline{x})q(\theta)d\theta$ กรณีที่ θ ต่อเนื่อง

$= \sum_{\theta} L(\underline{\theta}|\underline{x})q(\theta)$ กรณีที่ θ ไม่ต่อเนื่อง

เพื่อให้เข้าใจง่ายขึ้นจะแสดงตัวอย่างการใช้ฟังก์ชันก่อกวนการแจกแจงก่อน สมมติ นักฟิสิกส์ A และ B ทำการวัดค่าทางกายภาพค่าหนึ่งโดยที่ค่าวัดดังกล่าวแทนด้วย θ นักฟิสิกส์ A มีประสบการณ์ค่อนข้างสูงและทำการศึกษาในเรื่องนี้มามาก การประมาณ θ ของนักฟิสิกส์ A ดังกล่าวจึงมีความผิดพลาดน้อย ความเห็นของนักฟิสิกส์ A เกี่ยวกับ θ ก่อนเก็บข้อมูล สามารถประมาณด้วยการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ 900 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 20 นั่นคือ

$$p_A(\theta) = \frac{1}{\sqrt{2\pi}\,20}\exp\left[-\frac{1}{2}\left(\frac{\theta-900}{20}\right)^2\right]$$

ดังนั้นการแจกแจงก่อนของนักฟิสิกส์ A ที่มีเกี่ยวกับ θ คือ $\theta \sim N(900, 20^2)$ ส่วนนักฟิสิกส์ B นั้น มีประสบการณ์และความเชี่ยวชาญในเรื่องที่วัดน้อยกว่านักฟิสิกส์ A ความเห็นเกี่ยวกับ θ ของนักฟิสิกส์ B ก่อนเก็บข้อมูลสามารถประมาณด้วยการแจกแจง $N(800, 80^2)$ นั่นคือ

$$p_B(\theta) = \frac{1}{\sqrt{2\pi}\,80}\exp\left[-\frac{1}{2}\left(\frac{\theta-800}{80}\right)^2\right]$$

การแจกแจงก่อนของ θ ที่ได้จากนักฟิสิกส์ B จะมีค่าเฉลี่ยเท่ากับ 800 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 80 ซึ่งการแจกแจงของนักฟิสิกส์ A และนักฟิสิกส์ B แสดงได้ดังรูปที่ 5

ความคิดเห็นส่วนบุคคลหรือความเห็นก่อนการวัด θ ที่แตกต่างกันจะส่งผลให้การแจกแจงภายหลังของ θ แตกต่างกัน ถ้าให้ y เป็นค่าสังเกตที่ได้จากการวัดค่ากายภาพโดยที่ y มีการแจกแจงแบบปกติที่ค่าเฉลี่ย คือ θ และส่วนเบี่ยงเบนมาตรฐาน คือ σ นั่นคือ $f(y|\theta) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right]$ จะใช้ทฤษฎีเบย์ แสดงว่าความเห็นหลังจากเก็บข้อมูลเกี่ยวกับ θ มาแล้วหรือการแจกแจงภายหลังของ θ เมื่อกำหนด y มีความอ่อนไหวจากความเห็นก่อนของนักฟิสิกส์ทั้ง 2 ท่าน กล่าวคือ แต่ละการแจกแจงภายหลังของ θ หรือความ

เห็นหลังจากทำการวัดค่า θ มาแล้วจากนักฟิสิกส์ทั้ง 2 ท่านที่มีความเห็นเกี่ยวกับ θ ที่ต่างกันโดยผ่านวิธีการแบบเบย์ในการปรับค่า θ ใหม่จากข้อมูล y ที่วัดได้

สมมติว่าการแจกแจงก่อนหรือความคิดเห็นก่อนเกี่ยวกับ θ สามารถแทนได้ด้วยการแจกแจงแบบปกติ นั่นคือ $\theta \sim N(\theta_0, \sigma_0^2)$ และมีฟังก์ชันภาวะน่าจะของ θ คือ การแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ y และส่วนเบี่ยงเบนมาตรฐานเท่ากับ σ นั่นคือ

$$p(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right] \quad , \quad -\infty<\theta<\infty \qquad (2.16)$$

และ $$f(y|\theta) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right] \quad , \quad -\infty<y<\infty \qquad (2.17)$$

จาก (2.17) จะได้ $L(\theta|y) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{\theta-y}{\sigma}\right)^2\right]$

ดังนั้น การแจกแจงภายหลังของ θ เมื่อกำหนด y คือ

$$p(\theta|y) = \frac{L(\theta|y)p(\theta)}{\int_{-\infty}^{\infty} L(\theta|y)p(\theta)\,d\theta}$$

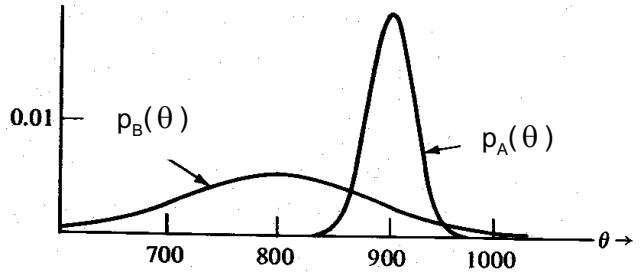
$$= \frac{\overline{\sigma^2}}{\sqrt{2\pi}}\exp\left[-\frac{1}{2\overline{\sigma^2}}(\theta-\bar{\theta})^2\right] \quad , \quad -\infty<\theta<\infty \qquad (2.18)$$

โดยที่ $\bar{\theta} = \frac{1}{\sigma_0^{-2}+\sigma^{-2}}\left(\sigma_0^{-2}\theta_0+\sigma^{-2}y\right)$ และ $\overline{\sigma^2} = \frac{1}{\sigma_0^{-2}+\sigma^{-2}}$ (การพิสูจน์อยู่ในภาคผนวก ก) นั่นคือ $\theta|y$ จะมีการแจกแจงแบบปกติค่าเฉลี่ย $\bar{\theta}$ และส่วนเบี่ยงเบนมาตรฐาน $\bar{\sigma}$ เมื่อค่าเฉลี่ยภายหลัง (Posterior mean : $\bar{\theta}$) คือ ค่าเฉลี่ยที่ถ่วงน้ำหนักด้วยค่าเฉลี่ยก่อน (Prior mean : θ_0) และค่าสังเกต y จากตัวอย่าง ถ้าให้ค่าวัดทางกายภาพ θ เมื่อกำหนด y มีการแจกแจงแบบปกติ คือ $\theta|y \sim N(850, 40^2)$ และ $L(\theta|y)$ สามารถแสดงได้ดังรูปที่ 6 ความเห็นภายหลังของค่ากายภาพของนักฟิสิกส์ A คือ $p_A(\theta|y)$ เป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ย 890 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 17.9 ส่วนความคิดเห็นของ B เมื่อทราบค่าสังเกต y คือ $p_B(\theta|y)$ จะมีการแจกแจงแบบปกติที่ค่าเฉลี่ย 840 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 35.7 ซึ่งการแจกแจงภายหลังทั้งสองเมื่อทราบค่าสังเกต y แสดงได้ดังรูปที่ 7 และตารางที่ 2 จะเห็นว่าแนวคิดหรือความเห็นของนักฟิสิกส์ A และนักฟิสิกส์ B ส่งผลต่อการพยากรณ์ค่ากาย

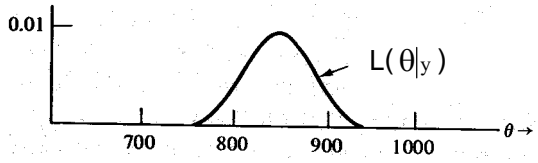
ภาพของ θ เมื่อทราบค่าสังเกต y ในตัวอย่างเดียวกันนี้ หากพิจารณาว่าไม่ใช่กรณีของการก่อกวนที่เกิดจากฟังก์ชันก่อกวนการแจกแจงก่อน อาจถือว่าเป็นความเห็นที่ต่างกันของนักฟิสิกส์ A และ B เกี่ยวกับ θ ก่อนการเก็บข้อมูล ซึ่งอาจนำทั้ง 2 ความเห็นนี้มาปรับความเห็นหลังจากเก็บข้อมูลโดยผ่านการอนุมานแบบเบย์ก็ได้

ตารางที่ 2 การแจกแจงก่อนและภายหลังของ θ จากนักฟิสิกส์ A และนักฟิสิกส์ B

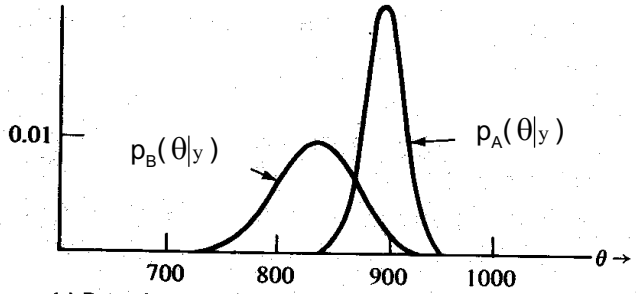
การแจกแจงก่อน	ภาวะน่าจะเป็นจากข้อมูล	การแจกแจงภายหลัง
นักฟิสิกส์ A $\theta \sim N(900,20^2)$	$N(850,40^2)$	นักฟิสิกส์ A $\theta \sim N(890,17.9^2)$
นักฟิสิกส์ B $\theta \sim N(800,80^2)$		นักฟิสิกส์ B $\theta \sim N(840,35.7^2)$



รูปที่ 5 การแจกแจงก่อนของนักฟิสิกส์ A และนักฟิสิกส์ B



รูปที่ 6 ฟังก์ชันภาวะน่าจะเป็นของ θ เมื่อกำหนด y = 850



รูปที่ 7 การแจกแจงภายหลังของนักฟิสิกส์ A และฟิสิกส์ B เมื่อกำหนด y = 850

2. ฟังก์ชันการก่อกวนโดยการตัดค่า (Deleted perturbution function) เป็นฟังก์ชันการก่อกวนที่ได้จากตัดค่าสังเกตตัวที่ i ออกจากชุดข้อมูล Kass , Tierney and Kadane (1989) นิยามฟังก์ชันการก่อกวนโดยการตัดค่า คือ $h_2^*(\theta) = [f(x_i|\theta)]^{-1}$ เมื่อ $f(x_i|\theta)$ คือ ฟังก์ชันความน่าจะเป็นของ x_i เมื่อกำหนด θ เพื่อให้ง่ายต่อการอธิบายจะเขียนฟังก์ชันการก่อกวนโดยการตัดค่าใหม่ เป็น $h_2^*(\theta) = [L(\theta|x_i)]^{-1}$ ฟังก์ชันการก่อกวนแบบนี้จะใช้วัดอิทธิพลของค่าสังเกตค่าที่ i ที่ถูกตัดออกจากชุดข้อมูล รูปแบบทั่วไปของฟังก์ชันความน่าจะเป็นภายหลังภายใต้ตัวแบบที่สมบูรณ์หรือตัวแบบเต็ม คือ M_0 : $p(\theta|\underline{x})$ ส่วนฟังก์ชันความน่าจะเป็นภายหลังภายใต้ตัวแบบที่ถูกก่อกวนหรือตัวแบบที่ถูกตัดค่าสังเกต x_i ออก คือ M_1 : $p(\theta|x^i)$ เมื่อ x^i คือ ชุดข้อมูลที่ค่าสังเกตตัวที่ i ถูกตัดออกไป นั่นคือ $x^i = \underline{x} - x_i \quad , i = 1,2,...,n$ ตัวแบบที่ถูกก่อกวนนี้เกิดจากการนำฟังก์ชันการก่อกวนโดยการตัดค่าไปก่อกวนตัวแบบที่สมบูรณ์ แสดงได้ดังนี้

พิจารณาตัวแบบสมบูรณ์จาก (2.14) จะได้ M_0 : $p(\theta|\underline{x}) = \frac{L(\theta|\underline{x})p(\theta)}{p(\underline{x})}$

นำฟังก์ชันการก่อกวนโดยการตัดค่า $h_2^*(\theta) = [L(\theta|x_i)]^{-1}$ ไปก่อกวนตัวแบบสมบูรณ์ จะได้ตัวแบบที่ถูกก่อกวน คือ

$$M_1 : p_1(\theta|x^i) = p(\theta|\underline{x})h_2^*(\theta) \qquad (2.19)$$

$$= \frac{L(\theta|\underline{x})p(\theta)h_2^*(\theta)}{p_2(x^i)}$$

$$= \frac{L(\theta|x_1)\text{K}\,L(\theta|x_i)\text{K}\,L(\theta|x_n)p(\theta)\,[L(\theta|x_i)]^{-1}}{p_2(x^i)}$$

$$= \frac{L(\theta|x_1)\text{K}\,L(\theta|x_{i-1})\,L(\theta|x_{i+1})\,\text{K}\,L(\theta|x_n)p(\theta)}{p_2(x^i)}$$

$$p_1(\theta|x^i) = \frac{L(\theta|x^i)p(\theta)}{p_2(x^i)} \qquad (2.20)$$

เมื่อ $p_2(x^i) = \int_\theta L(\theta|x^i)p(\theta)d\theta$ กรณีที่ θ ต่อเนื่อง

$= \sum_\theta L(\theta|x^i)p(\theta)$ กรณีที่ θ ไม่ต่อเนื่อง

การวิจัยครั้งนี้สนใจข้อมูลทั่วไปที่มีการแจกแจงแบบปกติ กำหนดเซตข้อมูลเริ่มต้นหรือข้อมูลที่สมบูรณ์ (Complete data) คือ $\underline{X} = \{X_1,...,X_n\}$ โดยที่ $X_i \sim N(\mu,\sigma^2)$, i = 1,2,...,n ในเบื้องต้นกล่าวไปแล้วว่าการแจกแจงก่อนที่ใช้คือ การแจกแจงก่อนแบบไม่มีสารสนเทศและไม่ทราบค่าพารามิเตอร์ μ และ σ^2 ดังนั้นจาก (2.13) จะได้ตัวแบบสมบูรณ์หรือตัวแบบเต็ม คือ

$$M_0 : p(\mu,\sigma|\underline{x}) = \sqrt{\frac{n}{2\pi}}\left[\frac{1}{2}\Gamma\left(\frac{n-1}{2}\right)\right]^{-1}\left(\frac{(n-1)s^2}{2}\right)^{\frac{n-1}{2}}\sigma^{-(n+1)}e^{-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\mu-\bar{x})^2\right]} \qquad (2.21)$$

และฟังก์ชันการก่อกวนโดยการตัดค่าสังเกต x_i ออก เมื่อ $x_i \sim N(\mu,\sigma^2)$, i = 1,2,...,n คือ

$$h_2^*(\mu,\sigma) = \left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\right\}^{-1} \quad ,-\infty<\mu<\infty \quad ,\sigma>0 \qquad (2.22)$$

นำฟังก์ชัน (2.22) ไปก่อกวนตัวแบบสมบูรณ์ จะได้ตัวแบบที่ถูกก่อกวนโดยฟังก์ชันการก่อกวนจากการตัดค่าสังเกต x_i ออก คือ

$$M_1:\ p(\mu,\sigma \mid x^i) = \sqrt{\frac{n-1}{2\pi}}\left[\frac{1}{2}\Gamma\left(\frac{n-2}{2}\right)\right]^{-1}\left(\frac{(n-2)s_i^{\,2}}{2}\right)^{\frac{n-2}{2}}\sigma^{-n}\,e^{-\frac{1}{2\sigma^2}\left[(n-2)s_i^{\,2}+(n-1)(\mu-\bar{x}_i)^2\right]} \quad (2.23)$$

เมื่อ x^i คือ เซตข้อมูลที่ตัด x_i ออกจากข้อมูล นั่นคือ $x^i = \{x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n\}$

$\bar{x}_i$ คือ ค่าเฉลี่ยที่เกิดจากข้อมูล x^i

s_i คือ ส่วนเบี่ยงเบนมาตรฐานที่เกิดจากข้อมูล x^i

ซึ่งจะหาอิทธิพลของค่าสังเกต x_i แต่ละตัวต่อไป เพื่อนำไปสู่การตัดสินใจว่าในชุดข้อมูลมีค่าผิดปกติปะปนมาหรือไม่ และค่าผิดปกติคือค่าสังเกตใด

นอกจากนี้ Weiss (1996) ได้เสนอรูปแบบของฟังก์ชันการก่อกวนอีก 2 ลักษณะที่ใช้ตรวจสอบความอ่อนไหวของตัวแบบเชิงเส้น โดยให้ m เป็นตัวแปรอิสระและ z เป็นตัวแปรตาม ฟังก์ชันแรก คือ ฟังก์ชันก่อกวนที่ใช้ตรวจความสอบอ่อนไหวต่อตัวแปรตาม (Sensitivity to z_i - values) สามารถประเมินได้จากฟังก์ชันก่อกวน $h_3^*(\theta,\delta) = \left\{\frac{f(z_i+\delta \mid \theta, m_i)}{f(z_i \mid \theta, m_i)}\right\}$ และฟังก์ชันก่อกวนที่ใช้ตรวจความสอบอ่อนไหวต่อตัวแปรอิสระ (Sensitivity to m_i - values) จะสามารถประเมินได้จากฟังก์ชันก่อกวน $h_4^*(\theta,\delta) = \left\{\frac{f(z_i \mid \theta, m_i+\delta)}{f(z_i \mid \theta, m_i)}\right\}$ เมื่อ δ คือ ค่าที่ต้องการก่อกวน

ตัวสถิติที่ใช้วัดความอ่อนไหว (Statistic for Measure Sensitivity)

ในกรณีใช้ฟังก์ชันการก่อกวนโดยการตัดค่าหรือการตัดค่าสังเกต x_i ออกจากข้อมูล Geisser (1980) เสนอให้ใช้ Conditional Predictive Ordinate (CPO) ซึ่งเป็นการวิเคราะห์ที่ใช้หลักการของเบย์เพื่อตรวจสอบค่าผิดปกติหรือค่าที่มีอิทธิพล โดยที่ CPO $= \left[E\left(h_2^*(\theta)\right)\right]^{-1}$ เมื่อ E(.) คือ ค่าคาดหวังที่ใช้เทียบกับการแจกแจงภายหลังของตัวแบบสมบูรณ์ สำหรับการประเมินว่าค่าสังเกตใดควรจะเป็นค่าผิดปกติจะพิจารณาจากค่า CPO ที่มีค่าต่ำที่สุดในชุดข้อมูลตามลำดับ

Pettit (1990) เสนอการใช้ตัวสถิติ CPO ในการตรวจสอบค่าผิดปกติในข้อมูลที่มีการแจกแจงแบบปกติ ดังนี้

1. ข้อมูลที่มีการแจกแจงแบบปกติตัวแปรเดียว (Univariate normal distribution data) กำหนด $x_1, x_2, ..., x_n$ เป็นข้อมูลที่มีการแจกแจงแบบ $N(\theta, \sigma^2)$ และการแจกแจงก่อนของ θ คือ $N(\mu, k\sigma^2)$ เมื่อกำหนด k และ σ^2 ค่าผิดปกติจะพิจารณาจาก CPO $= \left(n\bar{x} + \mu^{-1}\right)\left(n + k^{-1}\right)^{-1}$ ที่มีค่าต่ำที่สุดตามลำดับ ถ้าไม่ทราบค่า σ^2 แต่ทราบการแจกแจงก่อนของ σ^2 คือ การแจกแจงแบบอินเวอร์สไคร์สแคร์ (Inverted χ^2 prior distribution) ค่าผิดปกติจะพิจารณาจากค่าเฉลี่ยที่ได้ที่มีขนาดใหญ่ของการแจกแจงภายหลัง

2. ข้อมูลที่มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate normal distribution data) กำหนด $y_1, y_2, \ldots, y_n$ เป็นข้อมูลตัวอย่างที่สุ่มจากการแจกแจง $N_p(\theta^*, \sum^*)$ และการแจกแจงก่อนของ θ^* คือ $N_p(\mu^*, a^{-1}\sum^*)$ ถ้าทราบค่า $\sum^*$ สามารถตรวจสอบค่าผิดปกติจาก

$$\text{CPO} = \left(y_i - \frac{\sum_{i=1}^{n} y_i + a\mu}{n+a}\right)^T \frac{n+a}{n+a-1}\sum^{*-1}\left(y_i - \frac{\sum_{i=1}^{n} y_i + a\mu}{n+a}\right)$$

ที่มีค่าต่ำที่สุดตามลำดับ แต่ถ้า $\sum^*$ ไม่ทราบค่าและการแจกแจงก่อนของ $\sum^*$ เป็นการแจกแจงแบบ Inverted Wishart ซึ่งมีพารามิเตอร์ q และ Q (Box and Tiao 1973) ค่าผิดปกติจะพิจารณาจาก CPO ที่มีค่าต่ำที่สุดตามลำดับ ซึ่ง $\text{CPO} = \det(Q')$ เมื่อ $\det(\ )$ เป็นตัวกำหนด (Determinants) ของเมตริกซ์ และ Q'_i เป็นเมตริกซ์ขนาด $p \times p$ ที่

$$Q'_i = Q + \sum_{i \neq j} y_j y_j^T + a\mu\mu^T + \left(\sum_{i \neq j} y_j + a\mu\right)(n+a-1)^{-1}\left(\sum_{i \neq j} y_j + a\mu\right)^T$$

นอกจากนี้ Pettit (1990) ได้ใช้ CPO ตรวจสอบค่าผิดปกติในการวิเคราะห์การถดถอย และเสนอตัวสถิติ Ratio Ordinate Measure (ROM) โดยที่ ROM เกิดจากอัตราส่วนของ CPO กับค่าที่มากที่สุดของ CPO จากนั้นได้เปรียบเทียบ CPO และ ROM ในการตรวจสอบค่าผิดปกติในการวิเคราะห์การถดถอยทั้งในกรณีที่ทราบความแปรปรวน และกรณีที่ไม่ทราบความแปรปรวน

Pettit (1992) ศึกษาตัวประกอบเบย์ในตัวแบบค่าผิดปกติ (Outlier model) โดยใช้วิธีค่าสังเกตจินตภาพ (Device of imaginary observations) โดยการจำลองแบบข้อมูลจากฟังก์ชันการแจงแจงความน่าจะเป็น คือ f(x) และปะปนค่าผิดปกติลงใน f(x) จะได้ฟังก์ชันการแจกแจงความน่าจะเป็นที่ถูกปะปน คือ g(x) ในการจำลองแบบจะสมมติ f(x) เป็นฟังก์ชันความน่าจะเป็นที่ขึ้นกับพารามิเตอร์ θ ดังนั้น g(x) จะเป็น f(x) ที่ถูกปะปนค่าผิดปกติที่มีพารามิเตอร์เป็น $\theta+\delta$ หรือ $\delta\theta$ เมื่อ δ คือ ค่าผิดปกติที่ปะปน

การคำนวณตัวประกอบเบย์เพื่อตรวจสอบค่าผิดปกติ Pettit (1992) ทำการตรวจสอบกับข้อมูลตัวอย่างที่มีการแจกแจงแบบปกติตัวแปรเดียว การถดถอยเชิงเส้นอย่างง่าย (Simple linnear regression) และข้อมูลตัวอย่างแบบเอกซ์โพเนนท์เชียล (Exponential samples) สำหรับกรณีข้อมูลตัวอย่างที่มีการแจกแจงแบบปกติตัวแปรเดียว Pettit (1992) กำหนด M_a และ M_b เป็นตัวแบบที่ไม่มีค่าผิดปกติและตัวแบบที่มีค่าผิดปกติ 1 ค่า ตามลำดับ โดยที่ $M_a : y_1,y_2,\ldots,y_n \sim N(\mu,\sigma^2)$ และ $M_b : y_1,y_2,\ldots,y_{n-1} \sim N(\mu,\sigma^2)$, $y_n \sim N(\mu+\delta,\sigma^2)$ นั่นคือสมมติ y_n เป็นค่าผิดปกติ ตัวประกอบเบย์จากตัวแบบ M_a และ M_b คือ B_{01} โดยที่

$$B_{01} = \left\{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n-1}(y_i-y^*)^2}\right\}^{-\frac{n}{2}} \quad \text{เมื่อ} \quad \bar{y} = n^{-1}\sum_{i=1}^{n} y_i \quad \text{และ}$$

$$y^* = (n-1)^{-1}\sum_{i=1}^{n-1} y_i$$

ในการอภิปรายของ Pettit (1988) กล่าวว่า อาจเป็นไปได้ที่เกิดกรณีตรวจสอบข้อมูลแล้วไม่พบค่าผิดปกติทั้งที่มีค่าผิดปกติในข้อมูลนั้น หรือพบค่าผิดปกติแต่ไม่พบทั้งหมด (Masking effect) และอาจเป็นไปได้ที่เกิดกรณีตรวจสอบข้อมูลแล้วพบค่าผิดปกติทั้งที่ไม่มีค่าผิดปกติในข้อมูลนั้น หรือค่าที่ตรวจพบว่าเป็นค่าผิดปกติ ไม่ได้เป็นค่าผิดปกติ (Swamping effect) Pettit (1992) ได้พิจารณาในกรณีที่ค่าผิดปกติมี 2 ค่า โดยกำหนด M_c เป็นตัวแบบที่มีค่าผิดปกติ 2 ค่า โดยที่ $M_c : y_1, y_2, \ldots, y_{n-2} \sim N(\mu,\sigma^2)$, $y_{n-1} \sim N(\mu+\delta_1,\sigma^2)$, $y_n \sim N(\mu+\delta_2, \sigma^2)$ ตัวประกอบเบย์จากตัวแบบ M_a และ M_c คือ B_{02} โดยที่

$$B_{02} = \left\{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n-2}(y_i-y^{**})^2}\right\}^{-\frac{n}{2}} \quad \text{เมื่อ} \quad y^{**} = (n-1)^{-1}\sum_{i=1}^{n-2} y_i$$

นอกจากนี้ Pettit (1992) ได้อภิปรายผลการศึกษาตัวประกอบเบย์ กรณีข้อมูลตัวอย่างที่มีการแจกแจงแบบปกติตัวแปรเดียว ด้วยข้อมูลของ Freeman และข้อมูลของ Darwin กรณีข้อมูลการถดถอยเชิงเส้นอย่างง่าย ด้วยข้อมูลของ Mickey , Dunn and Clark และกรณีข้อมูลตัวอย่างแบบเอกซ์โพเนนท์เชียล ด้วยข้อมูลของ Kimber and Stevens

Geisser (1993) เสนอวิธีการตรวจสอบค่าผิดปกติที่ใช้ตัวประกอบเบย์เป็นพื้นฐานผ่านการทดสอบนัยสำคัญการพยากรณ์ สำหรับความไม่ลงรอย (Predictive signficance testing for discordancy) อาทิ (1) วัดการเบี่ยงเบน Δ_i (Deviation) ของ X_i จากการพยากรณ์ (Predictive) ของ ค่าเฉลี่ย (Mean) มัธยฐาน (Median) หรือฐานนิยม (Mode) (2) วัดความแตกต่างระหว่างการแจกแจงพยากรณ์ (Predictive distribution) ของตัวแบบสมบูรณ์และตัวแบบที่ตัดค่าสังเกต i ออก และ (3) ใช้ตัวสถิติ Condition Predictive Ordinate (CPO) และ Uncondition Predictive Ordinate (UPC)

Weiss (1996) ได้แสดงความสัมพันธ์ของตัวประกอบเบย์ และ Condition Predictive Ordinate (CPO) ตัวประกอบเบย์จะแทนด้วยสัญลักษณ์ $B(M_0, M_1)$ หมายถึง อัตราส่วนระหว่างการแจกแจงพยากรณ์ของตัวแบบสมบูรณ์ (M_0) และการแจกแจงพยากรณ์ของตัวแบบที่ถูกก่อกวน (M_1) ด้วยฟังก์ชันการก่อกวน $h_2^*(\theta)$ แสดงได้ดังนี้

$$
\begin{aligned}
B(M_0, M_1) &= \frac{f(\underline{x}|M_0)}{f(\underline{x}|M_1)} \\
&= \frac{\int f(\underline{x}|\theta)p(\theta)\,d\theta}{\int f(\underline{x}|\theta)p(\theta)h_2^*(\theta)\,d\theta} \\
&= \frac{\int f(\underline{x},\theta)\,d\theta}{\int f(\underline{x},\theta)h_2^*(\theta)\,d\theta} \\
&= \frac{f(\underline{x})}{\int f(\underline{x},\theta)h_2^*(\theta)\,d\theta \times \left(\frac{f(\underline{x})}{f(\underline{x})}\right)} \\
&= \frac{1}{\int f(\theta|\underline{x})h_2^*(\theta)\,d\theta}
\end{aligned}
$$

ดังนั้น $B(M_0, M_1) = \left[E\left(h_2^*(\theta)\,|\,\underline{x}\right)\right]^{-1}$ (2.24)

ซึ่งจะเห็นว่า CPO = $B(M_0, M_1)$

Young and Pettit (1996) ได้เสนอวิธีการวัดความไม่ลงรอยกัน (Measuring discordancy) ระหว่างการแจกแจงก่อนและข้อมูลค่าสังเกต สาเหตุในการวัดความไม่ลงรอยกันนั้นเพื่อต้องการตรวจสอบ หลังจากได้เห็นข้อมูลว่าความเห็นก่อนทำการเก็บรวบรวมข้อมูลของผู้เชี่ยวชาญหรือกลุ่มผู้เชี่ยวชาญนั้นถูกต้องและสอดคล้องกับข้อมูลหรือไม่ วิธีการพิจารณาภาวะที่ไม่ลงรอยกันนั้นจะใช้การวิเคราะห์ตัวประกอบเบย์ (Bayes factor analysis) โดยใช้การแจกแจงมาร์จินัลของข้อมูลซึ่งได้จาก Proper prior distribution เปรียบเทียบกับการใช้การแจกแจงมาร์จินัลของข้อมูลโดยที่การแจกแจงก่อนเป็นแบบไม่มีสารสนเทศ พิจารณาข้อมูลตัวอย่างขนาด n โดยที่ค่าสังเกตที่ได้จากการสุ่มจากการแจกแจงแบบปกติ $N(\theta, \sigma^2)$ ซึ่ง θ ไม่ทราบค่าแต่ σ^2 ทราบค่า และมีการแจกแจงก่อนของ θ คือ $N(\theta_0, \sigma_0^2)$ ซึ่ง θ_0 และ σ_0^2 ทราบค่า และสมมติต่อไปว่าไม่มีสารสนเทศก่อนสำหรับ θ และให้ $p(\theta) = c\left(2\pi\sigma^2\right)^{-\frac{1}{2}}$ ดังนั้นทำให้ได้ตัวประกอบเบย์ (B) เท่ากับ $\sqrt{\frac{n}{c(nk+1)}}\exp\left\{\frac{-n(\bar{y}-\theta_0)^2}{2\sigma^2(nk+1)}\right\}$ ถ้า B = 0 แสดงว่ามีความไม่ลงรอยกันระหว่างการแจกแจงก่อนกับข้อมูลสูง แต่ในการวิเคราะห์ตัวประกอบเบย์นี้ไม่ใช่การวัดอิทธิพลของการแจกแจงก่อนบนการแจกแจงภายหลังเป็นเพียงการวัดความขัดแย้งระหว่างการแจกแจงก่อนกับข้อมูลที่ถูกเก็บรวบรวมมา

การประเมินอิทธิพลของตัวก่อกวน (Influence Assessment of Perturbution)

ในการประเมินอิทธิพลของตัวก่อกวน ในฟังก์ชันการก่อกวน $h^*(\theta)$ สามารถประเมินโดยค่าวัดการเบี่ยงเบน (Divergence measure) ระหว่างการแจกแจงภายหลังของตัวแบบที่ถูกก่อกวนและการแจกแจงภายหลังของตัวแบบสมบูรณ์ กำหนดให้ $\tau(\theta)$ คือ ฟังก์ชันของ θ ที่เป็นฟังก์ชันที่ต้องการประเมินอิทธิพลในการวิเคราะห์และนิยามค่าวัดการเบี่ยงเบนดังนี้

$$D_{\tau(\theta)}\left(g, h^*\right) = \int g\left[\frac{p_1\left(\tau(\theta)\middle|\underline{x}\right)}{p\left(\tau(\theta)\middle|\underline{x}\right)}\right] p\left(\tau(\theta)\middle|\underline{x}\right) d\theta \qquad (2.25)$$

โดยที่ g(.) เป็นฟังก์ชัน Convex และ g(1) เท่ากับ 0 เพราะฟังก์ชัน g เป็นฟังก์ชันที่ใช้วัดอิทธิพลจากฟังก์ชันการก่อกวน ถ้า $\left\{p_1\left(\tau(\theta)\middle|\underline{x}\right)/p\left(\tau(\theta)\middle|\underline{x}\right)\right\}$ เท่ากับ 1 แสดงว่าการก่อกวนของฟังก์ชันการก่อกวนที่อยู่ในตัวแบบการที่ถูกก่อกวนไม่ส่งผลกระทบต่อตัวแบบสมบูรณ์ หรือหมายความว่าตัวแบบการที่ถูกก่อกวนและตัวแบบสมบูรณ์ไม่มีความต่างกัน ซึ่งมีนักสถิติหลายท่านได้นำเสนอฟังก์ชัน g(.) นี้เอาไว้ดังที่จะกล่าวต่อไป

การแจกแจงภายหลังที่ถูกก่อกวนด้วยฟังก์ชันการก่อกวน h^* (Posterior perturbation distribution) เกิดจากการคูณ $p(\theta|\underline{x})$ ด้วย $h^*(\theta)$ ดังนั้น $p_1(\theta|\underline{x}) = \left\{p(\theta|\underline{x})h^*(\theta)/E\left[h^*(\theta)|\underline{x}\right]\right\}$ จากสมการนี้ทำให้ได้ทฤษฎีที่ 2

ทฤษฎีที่ 2 ทฤษฎีการแจกแจงมาร์จินัลเบย์ (Marginal bayes distribution theorem) (Weiss 1996)

$$\frac{p_1(\tau|\underline{x})}{p(\tau|\underline{x})} = \frac{E\left[h^*\{\underline{\theta}(\tau;\rho)\}|\underline{x},\tau\right]}{E\left[h^*(\underline{\theta})|\underline{x}\right]} \equiv h^*(\tau)$$

เมื่อ ρ ถูกเลือกเพื่อทำให้ (τ , ρ) เป็นการแปลงแบบ 1-1 และเมเชอเรเบิลของ $\underline{\theta}$ และ $h^*(\tau)$ แตกต่างจาก $h^*(\underline{\theta})$ โดยอาร์กูเมนต์

เมื่อสนใจในพารามิเตอร์ τ มากกว่าเวกเตอร์พารามิเตอร์ $\underline{\theta}$ ทั้งหมด จากทฤษฎีนี้สรุปว่าถ้าสนใจเฉพาะ τ ฟังก์ชันการก่อกวนที่เหมาะสมน่าจะเป็น $h^*(\tau)$ มากกว่า $h^*(\underline{\theta})$ โดย τ ได้จากเทคนิคการแปลงตัวแปรจาก $\underline{\theta}$ เป็น (τ , ρ)

สำหรับฟังก์ชัน $g(\ .\ )$ ที่มีผู้เสนอไว้มีหลายฟังก์ชันด้วยกัน ซึ่งพอสรุปได้ดังนี้

1. ฟังก์ชัน $g_1(a) = -\log_e a$ ให้ไว้โดย Kullback จึงเรียกการวัดการเบี่ยงเบนแบบนี้ว่า การวัดการเบี่ยงเบนแบบคูลแบคก์ (Kullback divergence : D_1) โดยที่

$$D_1 = D_{\tau(\theta)}\left(g_1, h^*\right) = \int -\log_e\left[\frac{p_1\left(\tau(\theta)|x^i\right)}{p(\tau(\theta)|\underline{x})}\right]p(\tau(\theta)|\underline{x})d\theta \qquad (2.26)$$

คือ negative log-posterior geometric mean for $h^*(\tau)$

2. ฟังก์ชัน $g_2(a) = a\log_e a$ ให้ไว้โดย Bernardo การวัดการเบี่ยงเบนโดยฟังก์ชันนี้จะเรียกว่า การวัดการเบี่ยงเบนแบบเบอนาร์โด (Bernardo divergence : D_2) โดยที่

$$D_2 = D_{\tau(\theta)}\left(g_2, h^*\right) = \int\left\{\left[\frac{p_1\left(\tau(\theta)|x^i\right)}{p(\tau(\theta)|\underline{x})}\right]\log_e\left[\frac{p_1\left(\tau(\theta)|x^i\right)}{p(\tau(\theta)|\underline{x})}\right]\right\}p(\tau(\theta)|\underline{x})d\theta \qquad (2.27)$$

คือ log-perturbed-posterior geometric mean for $h^*(\tau)$

3. ฟังก์ชัน $g_3(a) = a\log_e a - \log_e a$ ฟังก์ชันนี้เกิดจากผลรวมของฟังก์ชัน $g_1(a)$ และ $g_2(a)$ ดังกล่าวข้างต้น สัญลักษณ์ที่ใช้สำหรับวัดการเบี่ยงเบน $g_3(a)$ คือ D_3 โดยที่

$$\begin{aligned} D_3 &= D_1 + D_2 \\ &= D_{\tau(\theta)}\left(g_3, h^*\right) \\ &= \int\left\{\left[\frac{p_1\left(\tau(\theta)\middle|x^i\right)}{p\left(\tau(\theta)\middle|\underline{x}\right)}\right]\log_e\left[\frac{p_1\left(\tau(\theta)\middle|x^i\right)}{p\left(\tau(\theta)\middle|\underline{x}\right)}\right]-\log_e\left[\frac{p_1\left(\tau(\theta)\middle|x^i\right)}{p\left(\tau(\theta)\middle|\underline{x}\right)}\right]\right\}p\left(\tau(\theta)\middle|\underline{x}\right)d\theta \end{aligned} \quad (2.28)$$

4. ฟังก์ชัน $g_4(a) = 0.5|a-1|$ สัญลักษณ์ที่ใช้ในการวัดค่าเบี่ยงเบน $g_4(a)$ คือ L_1 (L_1 – Divergence) ซึ่งต่อไปจะเรียกว่าตัวสถิติที่ใช้วัดอิทธิพลแบบระยะทาง L_1 (L_1 – distance influence statistic) โดยที่

$$L_1(h^*) = D_{\tau(\theta)}\left(g_4, h^*\right) = 0.5\int\left|\frac{p_1\left(\tau(\theta)\middle|x^i\right)}{p\left(\tau(\theta)\middle|\underline{x}\right)}-1\right|p\left(\tau(\theta)\middle|\underline{x}\right)d\theta \quad (2.29)$$

คือ posterior mean absolute

5. ฟังก์ชัน $g_5(a) = (a-1)^2$ เรียกการเบี่ยงเบนที่เกิดจากฟังก์ชันนี้ว่า ค่าเบี่ยงเบนแบบไคว์-สแควร์ (Chi-square divergence) สัญลักษณ์ที่ใช้สำหรับค่าเบี่ยงเบน $g_5(a)$ นี้คือ χ^2 (χ^2 - divergence statistic) โดยที่

$$\chi^2(h^*) = D_{\tau(\theta)}\left(g_5, h^*\right) = \int\left\{\frac{p_1\left(\tau(\theta)\middle|x^i\right)}{p\left(\tau(\theta)\middle|\underline{x}\right)}-1\right\}^2 p\left(\tau(\theta)\middle|\underline{x}\right)d\theta \quad (2.30)$$

คือ posterior variance ซึ่งสัมพันธ์กับตัวสถิติที่ใช้วัดอิทธิพลและความอ่อนไหวในการวิเคราะห์แบบเบย์ในงานวิจัยของ Kass , Tierney and Kadane (1989) คือ ตัวสถิติที่ใช้วัดการเปลี่ยนแปลงที่มีค่ามาตรฐานสูงสุด (Maximum Standardized Change Statistic) หรือตัวสถิติ MSC (MSC Statistic) ความสัมพันธ์ดังกล่าว คือ $MSC(h^*) = \left\{\chi^2(h^*)\right\}^{\frac{1}{2}}$

ทฤษฎีที่ 3 ที่จะกล่าวต่อไป เป็นทฤษฎีที่ใช้พิจารณาอิทธิพลเชิงพยากรณ์ (Predictive influence) เมื่อใช้ $h^*(\tau)$ เทียบกับ $h^*(\theta)$

ทฤษฎีที่ 3 ทฤษฎีความสัมพันธ์ระหว่างอิทธิพลของฟังก์ชันการก่อกวนบน θ และอิทธิพลของฟังก์ชันการก่อกวนบน $\tau(\theta)$ (Weiss and Cook 1992)

$$0 \le D_{\tau}\left(g,h^*\right) = D_{h^*(\tau)}\left(g,h^*\right) \le D_{\theta}\left(g,h^*\right) = D_{h^*(\theta)}\left(g,h^*\right) \qquad (2.31)$$

ทฤษฎีนี้ได้จากสมบัติ Convex ของฟังก์ชัน $g(.)$ และทฤษฎีของเจนเซ่น (Jensen's theorem) กล่าวคือ อิทธิพลที่เกิดจากการก่อกวนของฟังก์ชันการก่อกวนจะไม่มีค่าเป็นลบและอิทธิพลของการก่อกวนที่เกิดจากฟังก์ชันการก่อกวนบนการแจกแจงภายหลังของ τ จะมีค่าไม่มากกว่าอิทธิพลของการก่อกวนที่เกิดจากฟังก์ชันการก่อกวนบนการแจกแจงภายหลังของ θ

อุปสรรคสำคัญของตัววัดการเบี่ยงเบนที่ได้จากการประเมินความอ่อนไหวแบบเบย์ คือ การตีความหรือให้ความหมายของการเบี่ยงเบน มีหลายงานวิจัยที่พยายามจะตีความการเบี่ยงเบนให้เข้าใจได้ง่ายขึ้น สรุปได้ดังนี้ Johnson and Geisser (1983) และ Pettit and Smith (1985) ศึกษาความอ่อนไหวแบบเบย์ โดยประเมินค่าเบี่ยงเบนในเทอมของ Leverage และ Residual McCulloch (1989) เปรียบการวัดการเบี่ยงเบนแบบเบอนาร์โดว่าเป็นการวัดการเบี่ยงเบนระหว่างฟังก์ชันการแจกแจงความน่าจะเป็นของตัวแปรเดียว 2 ฟังก์ชัน และ Carlin and Polson (1991) เปรียบการวัดการเบี่ยงเบนแบบคูลแบคก์ ว่าเป็นค่าคาดหวังในตัวอย่างสุ่มเชิงพยากรณ์แบบทวนซ้ำ (Repeated predictive sampling) ความยากของการวัดการเบี่ยงเบนแบบคูลแบคก์ คือ ค่าที่ได้จากการคำนวณตัวสถิติที่วัดการเบี่ยงเบนแบบคูลแบคก์ ไม่มีจุดตัดสินที่แน่นอน ว่าการก่อกวนโดย h^* มีอิทธิพลต่อตัวแบบสมบูรณ์หรือไม่ ทั้งนี้ขึ้นอยู่กับข้อมูลแต่ละชุด

Weiss (1996) ได้เสนอให้ใช้ตัววัดการเบี่ยงเบนที่สามารถตีความได้ง่าย คือ การวัดการเบี่ยงเบนแบบ L_1 ตาม (2.29) โดยตัววัดการเบี่ยงเบนนี้เป็นตัวสถิติที่ใช้วัดระยะทางระหว่างความหนาแน่น 2 ความหนาแน่น ดังนั้น เพื่อให้สอดคล้องกับความเป็นมาของตัวสถิติจะเรียกตัวสถิตินี้ว่า ตัวสถิติวัดอิทธิพลแบบระยะทาง L_1 หรือตัวสถิติระยะทาง L_1(L_1- distance statistic)

จาก (2.29) นั่นคือ $L_1(h^*) = \frac{1}{2}\int \left| \frac{p_1\left(\tau(\theta)|\underline{x}\right)}{p\left(\tau(\theta)|\underline{x}\right)} - 1 \right| p\left(\tau(\theta)|\underline{x}\right) d\theta$ จะได้

$$L_1(h^*) = \frac{1}{2}\int \left| p_1\left(\tau(\theta)|\underline{x}\right) - p\left(\tau(\theta)|\underline{x}\right) \right| d\theta \qquad (2.32)$$

$L_1(h^*)$ มีขอบเขตจำกัด คือ มีค่าระหว่างศูนย์ถึงหนึ่ง ($0 \le L_1(h^*) \le 1$) ถ้า $L_1 = 0$ แสดงว่าการก่อกวนที่เกิดจากฟังก์ชันการก่อกวนไม่มีผลกระทบต่อตัวแบบสมบูรณ์หรือไม่มีความแตกต่างกัน ระหว่างการแจกแจงภายหลังของตัวแบบที่ถูกก่อกวนกับการแจกแจงภายหลังของตัวแบบสมบูรณ์ ในกรณีที่ฟังก์ชันการก่อกวนเป็นฟังก์ชันการก่อกวนแบบตัดค่าสังเกต x_i ออกจากข้อมูล ความหมายที่ได้ คือ ค่าสังเกตที่ i ที่ถูกตัดออกไม่เป็นค่าที่มีอิทธิพลหรือไม่เป็นค่าที่ผิดปกติในชุดข้อมูล แต่ถ้า $L_1 = 1$ แสดงว่ามีความแตกต่างกันอย่างสิ้นเชิงระหว่างการแจกแจงภายหลังของตัวแบบที่ถูกก่อกวนกับความหนาแน่นภายหลังของตัวแบบสมบูรณ์หรือการก่อกวนโดยฟังก์ชัน h^* ส่งผลกระทบต่อตัวแบบสมบูรณ์มาก ในกรณีที่ฟังก์ชันการก่อกวน คือ $h^*_2(\theta)$ จะหมายความว่า ค่าสังเกต x_i ที่ถูกตัดออกจากข้อมูลเป็นค่าสังเกตที่มีอิทธิพลสูงมากหรืออาจพิจารณาได้ว่าเป็นค่าที่ผิดปกติในชุดข้อมูล

การพิจารณาว่าค่าสังเกตใดเป็นค่าผิดปกติสำหรับตัวสถิติที่อิงแนวคิดแบบเบย์ จะพิจารณาจากค่าสถิติ ถ้าค่าสถิติของค่าสังเกตใดห่างจากค่าสถิติของค่าสังเกตอื่น ๆ มาก จะพิจารณาค่าสังเกตนั้นเป็นค่าผิดปกติ ทั้งนี้การพิจารณาว่าค่าสังเกตใดเป็นค่าผิดปกติหรือเป็นค่าที่มีอิทธิพลหรือไม่ ยังขึ้นกับข้อมูลในแต่ละชุด โดยไม่มีการตัดสินว่าต้องมากกว่าค่าใดค่าหนึ่ง ตารางที่ 3 เป็นการพิจารณาค่าผิดปกติของ Weiss (1996) ด้วยตัวสถิติ CPO ตัวสถิติเบอนาร์โด ตัวสถิติ L_1 และ ตัวสถิติ χ^2 จะเห็นว่าทั้ง 4 ตัวสถิติให้ข้อสรุปไปในแนวทางเดียวกัน คือ ข้อมูลตัวที่ 15 , 30 และ 58 เป็นค่าผิดปกติ ที่น่าสังเกต คือ ตัวสถิติ L_1 จะเห็นว่าระยะทางจากตัวสถิติของค่าสังเกตที่ 58 เท่ากับ 0.391 ซึ่งไม่ถือว่าเข้าใกล้ 1 แต่เมื่อเทียบกับค่าสถิติอื่น ๆ ภายในชุดข้อมูลแล้ว พบว่าระยะทางที่ได้ค่อนข้างห่างมาก ดังนั้น ค่าสังเกตที่ 58 จึงเป็นค่าผิดปกติ ซึ่งเป็นผลให้ค่าสังเกตที่ 15 และ 30 เป็นค่าผิดปกติด้วย

ตารางที่ 3 ตัวอย่างการพิจารณาค่าผิดปกติด้วยตัวสถิติที่อิงแนวคิดแบบเบย์

ข้อมูลลำดับที่	100*CPO	L_1	เบอนาร์โด	χ^2
15	0.0016	0.656	2.160	71
30	0.0200	0.633	1.970	122
58	0.0220	0.391	0.600	3.67
19	0.5870	0.122	0.066	0.20
56	0.6770	0.100	0.044	0.14
Mininmum value	6.53	0.049	0.008	0.017
Lower guartile	3.21	0.067	0.016	0.033
Median value	2.04	0.084	0.024	0.051
Upper quartile	1.04	0.136	0.066	0.165
Maximum value	0.0016	0.656	2.160	122

ทฤษฎี วิธีการและงานวิจัยที่เกี่ยวข้องกับตัวสถิติที่อิงแนวคิดแบบดั้งเดิม

การตรวจสอบค่าผิดปกติ 1 ค่าและ 2 ค่า (One and two outlier procedures)

วิธีการตรวจสอบค่าผิดปกติโดยตัวสถิติที่อิงแนวคิดแบบดั้งเดิม ได้มีผู้ศึกษาและพัฒนาตัวสถิติอย่างต่อเนื่อง ในช่วงศตวรรษที่ 20 นักสถิติหลายท่านได้พัฒนาวิธีการตรวจสอบค่าผิดปกติ โดยทำการตรวจสอบค่าผิดปกติในข้อมูลตัวอย่างที่สุ่มจากประชากรที่มีการแจกแจงแบบปกติ

Grubbs (1950) และ Dixon (1953) สร้างแบบจำลองเพื่อศึกษาการแจกแจงของข้อมูลตัวอย่างที่มีค่าผิดปกติในชุดข้อมูล โดยการจำลองแบบมี 2 รูปแบบ คือ รูปแบบที่ 1 เป็นรูปแบบที่ข้อมูลมีลักษณะการแจกแจงระหว่าง 2 การแจกแจงแบบปกติที่มีค่าเฉลี่ยต่างกัน แต่ความแปรปรวนเท่ากัน นั่นคือ $N(\mu,\sigma^2)$ และ $N(\mu+\omega,\sigma^2)$ และรูปแบบที่ 2 เป็นรูปแบบที่ข้อมูลมีลักษณะการแจกแจงระหว่างการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากัน แต่ความแปรปรวนต่างกัน นั่นคือ $N(\mu,\sigma^2)$ และ $N(\mu,(\omega\sigma)^2)$ ทั้งสองรูปแบบนี้ยังพบในงานวิจัยของ Murphy(1951) Paulson (1952) Ferguson (1961) Prescott (1979) Resner (1983) Brant (1990) Davies and Gather (1993) เป็นต้น

ในช่วงแรกของการพัฒนาตัวสถิติที่อิงแนวคิดแบบดั้งเดิม เป็นการศึกษาค่าผิดปกติที่เกิดขึ้น 1 ค่า โดยการพัฒนาหาวิธีการตรวจสอบค่าผิดปกติที่เป็นรูปแบบมีมากขึ้น ส่วนใหญ่นิยมใช้การจำลองแบบข้อมูลในการศึกษา ได้มีการพัฒนาเทคนิคและวิธีการต่าง ๆ เป็นลำดับกล่าวโดยสรุป คือ Thomson (1935) แสดงการแจกแจงของ y_i เมื่อ $y_i = \frac{x_i - \bar{x}}{s}$ โดยแสดงให้เห็นว่า $\sqrt{\frac{n-2}{n-1-y_i}}\,y_i$ มีการแจกแจงแบบที ที่องศาอิสระ n-1 ต่อมามีผู้ชี้ให้เห็นถึงปัญหาที่เกิดขึ้นเมื่อใช้ค่าวิกฤต และพบว่าค่าวิกฤตที่ Thomson (1935) ให้ไว้เหมาะสำหรับกรณีที่ $\text{Max}|y_i|$ และไม่ได้มีการแจกแจงที่แท้จริง (Exact distribution) ต่อมา Grubbs (1950) ได้ศึกษาการแจกแจงที่แท้จริงของ $\text{Max}|y_i|$ โดยใช้เทคนิคการอินทิเกรตแบบวิธีเชิงตัวเลข สำหรับการตรวจสอบค่าผิดปกติ 1 ค่า นอกจากนี้ ในการศึกษาค่าผิดปกติเพียง 1 ค่า Dixon (1953) เสนอการตรวจสอบค่าผิดปกติโดยใช้อัตราส่วนระหว่างค่าสังเกตที่สงสัยว่าจะเป็นค่าผิดปกติ กับค่าสังเกตที่อยู่ใกล้ที่สุดหรือค่าสังเกตที่อยู่ใกล้ที่สุดตัวถัดไปเทียบกับพิสัยของข้อมูล Paulson (1952) ใช้ตัวสถิติ Extreme Studentized Deviate (ESD) ในการตรวจสอบค่าผิดปกติ 1 ค่า โดยมองปัญหาค่าผิดปกติ เป็นกรณีเฉพาะของตัวอย่าง 2 ตัวอย่างที่มีการซ้อนทับกัน Ferguson (1961) เสนอให้ใช้ตัวสถิติ Skewness และ

Kurtosis ในการตรวจสอบค่าผิดปกติ 1 ค่า โดยที่ Skewness = $\frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}}$ และ

Kurtosis = $\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$

จากการวิจัยของนักสถิติหลายท่านพบว่า การตรวจสอบค่าผิดปกติเพียง 1 ค่า ไม่ครอบคลุมที่จะตอบคำถามทั้งหมด บางครั้งพบว่าในข้อมูลที่มีค่าผิดปกติ 2 ค่า แต่เมื่อตรวจสอบค่าผิดปกติในข้อมูลดังกล่าว ด้วยวิธีการที่ใช้ตรวจสอบค่าผิดปกติ 1 ค่าแล้วไม่พบค่าผิดปกติในชุดข้อมูล จึงมีผู้เสนอวิธีการและตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติที่ละ 2 ค่า เช่น Grubbs (1950) เสนอตัวสถิติ $\frac{S^2_{n,n-1}}{S}$, $\frac{S^2_{1,2}}{S}$ และ $\frac{S^2_{1,n}}{S}$ สำหรับใช้ตรวจสอบค่าผิดปกติ 2 ค่า ในลักษณะที่ต่างกันไป Murphy (1951) เสนอว่าผลรวมของส่วนเหลือมาตรฐานแบบสติวเดนท์ (Student residual) ที่มีค่าใหญ่สุด 2 ค่า ใช้ทดสอบค่าผิดปกติ 2 ค่า การศึกษาค่าผิดปกติทีละ 2 ค่า ยังมีนักสถิติที่ทำการศึกษาไว้ อาทิ Dixon (1953) McMillan (1971) Tietjen and Moore (1972) Rosner (1975) Prescott (1979) เป็นต้น

การตรวจสอบค่าผิดปกติทีละหลายค่า (Many outlier procedures)

ปัญหาหนึ่งที่พบในการตรวจสอบค่าผิดปกติ ก็คือ การที่ตรวจสอบค่าผิดปกติด้วยวิธีการตรวจสอบที่ละ 1 ค่าหรือ 2 ค่า แล้วไม่สามารถอธิบายได้ว่าแท้จริงแล้วในข้อมูลมีค่าผิดปกติทั้งหมดกี่ค่า ซึ่งอาจเป็นไปได้ว่ามีค่าผิดปกติมากกว่า 2 ค่า เมื่อนำตัวสถิติที่พัฒนาขึ้นไปใช้จริง นักสถิติส่วนใหญ่ที่ศึกษาตัวสถิติที่อิงแนวคิดดั้งเดิมต่างตระหนักถึงปัญหานี้ มีความพยายามพัฒนาวิธีการและตัวสถิติที่สามารถตรวจสอบค่าผิดปกติได้ที่ละหลายค่า สมมติ k ค่า โดยที่ $k \geq 2$

Tietjen and Moore (1972) พัฒนาตัวสถิติ Grubbs (1950) ให้สามารถตรวจสอบค่าผิดปกติได้หลายค่า (k ค่า) จากขนาดตัวอย่าง n โดยที่ $k \leq \frac{n}{2}$ ตัวสถิติดังกล่าว คือ ตัวสถิติ L_k และ E^*_k ผลการศึกษาพบว่าตัวสถิติ L_k และ E^*_k สามารถตรวจสอบค่าผิดปกติได้ k ค่า ที่ปรากฏอยู่ด้านใดด้านหนึ่งของของข้อมูล และลดปัญหาการแอบแฝงของค่าผิดปกติ

(Masking effect) ลงได้ นอกจากนี้ Tietjen and Moore (1972) ได้เปรียบเทียบอำนาจการทดสอบของตัวสถิติ L_k และ E^*_k กับตัวสถิติของ Dixon (1953) ตัวสถิติของ Grubbs (1950) และตัวสถิติของ Ferguson (1961) ผลการศึกษาพบว่าในการตรวจสอบค่าผิดปกติที่เกิดขึ้นมากกว่าหรือเท่ากับ 2 ค่า ตัวสถิติ L_k และ E^*_k มีอำนาจการทดสอบสูงกว่าตัวสถิติอื่น ๆ

Rosner (1975) ได้เปรียบเทียบวิธีการตรวจสอบค่าผิดปกติทีละ 1 ค่าและ 2 ค่า กับวิธีการตรวจสอบค่าทีละหลายค่า โดยใช้ตัวสถิติดังนี้

1.Extreme Studentized Deviate (ESD) โดยที่ $ESD = \dfrac{\text{Max}\left|x_i - \bar{x}\right|}{s}$

2.Studentized Range (STR) โดยที่ $STR = \dfrac{x_{[n]} - x_{[1]}}{s}$

เมื่อ $x_{[n]}$ และ $x_{[1]}$ คือ ข้อมูลที่ค่ามากที่สุดและน้อยที่สุดตามลำดับ

3.Kurtosis (KUR) โดยที่

$$KUR = \frac{n\sum\left(x_i - \bar{x}\right)^4}{\left[\sum\left(x_i - \bar{x}\right)^2\right]^2}$$

4.R-Statistic (RST) โดยที่ $RST = \dfrac{\text{Max}\left|x_i - a\right|}{b}$

เมื่อ $a = \dfrac{\sum_{i=k+1}^{n-k} x_{[i]}}{n-2k}$ = Trimmed mean

$b^2 = \dfrac{\sum_{i=k+1}^{n-k}\left(x_{[i]} - a\right)^2}{n-2k-1}$ = Trimmed variance

ในการศึกษา Rosner (1975) ได้หาค่าวิกฤตของตัวสถิติทั้ง 4 ตัว โดยใช้เทคนิคมอนติ คาร์โล ในการประมาณค่าวิกฤต ผลจากการศึกษาพบว่าการตรวจสอบค่าผิดปกติทีละหลายมีประสิทธิภาพสูงกว่าการตรวจสอบค่าผิดปกติทีละ 1 ค่าและ 2 ค่า ในกรณีที่ค่าผิดปกติมีมากกว่า 1 ค่า และตัวสถิติ ESD เป็นตัวสถิติที่มีอำนาจการทดสอบสูงที่สุดและสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าการทดสอบด้วยตัวสถิติอื่นอีก 3 ตัว ทั้งนี้เพราะตัวสถิติ ESD มีการคำนวณที่สมเหตุสมผล Rosner (1975) ให้ตารางค่าวิกฤตที่ระดับนัยสำคัญ 0.05 0.01 และ 0.005 สำหรับขนาดตัวอย่างเริ่มต้นจาก 10 จนถึง 50 และ k = 1 และ 2

Rosner (1983) ชี้ให้เห็นถึงปัญหาของวิธีการตรวจสอบค่าผิดปกติทีละหลายค่าโดยใช้ตัวสถิติ ESD ในการตรวจสอบค่าผิดปกติ กล่าวคือ เริ่มแรกของการพัฒนาวิธีการตรวจสอบค่าผิดปกติทีละหลายค่าโดยใช้ตัวสถิติ ESDได้กำหนดให้ค่าวิกฤตของตัวสถิติทั้งหมดอยู่ที่ระดับเดียวกันเพื่อความสะดวก ปัญหาจากการทำเช่นนี้ Hawkins (1978) ได้อภิปรายไว้ถึงความไม่เหมาะสม นั่นคือ วิธีการนี้จะมีความคลาดเคลื่อนประเภทที่ 1 ที่เหมาะสมในกรณีที่ไม่มีค่าผิดปกติในข้อมูล แต่ถ้ามีค่าผิดปกติในข้อมูลวิธีการนี้อาจให้ความคลาดเคลื่อนประเภทที่ 1 ที่ไม่เหมาะสม ต่อมา Rosner (1983) ได้พัฒนาตัวสถิติ ESD ดังที่จะได้กล่าวรายละเอียดต่อไป Brant (1990) เปรียบเทียบตัวสถิติของ Rosner (1983) และตัวสถิติของ Tukey (1977) ในการตรวจสอบค่าผิดปกติ Davies and Gather (1993) ศึกษาลักษณะเฉพาะของค่าผิดปกติหลายค่าโดยตัวสถิติที่ปรับปรุงจาก Rosner (1983) นอกจากนี้ ยังมีการประยุกต์ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม ไปใช้ในการตรวจสอบค่าผิดปกติสำหรับการวิเคราะห์การถดถอย ซึ่งจะไม่กล่าวรายละเอียดในที่นี่

ตัวสถิติ Generalized Extreme Studentized Deviate (GESD)

Rosner (1983) ได้พัฒนาวิธีการตรวจสอบค่าผิดปกติของ Rosner (1975) โดยการพัฒนานี้ทำให้มีแนวโน้มที่จะตรวจสอบค่าผิดปกติที่มีในข้อมูลได้ถูกต้องแม่นยำกว่าเดิม เพราะสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้เหมาะสมกว่า วิธีที่ว่านี้ คือ Generlized Extreme Studentized Deviate (GESD) ซึ่งมีพื้นฐานจากตัวสถิติ R_1 , R_2 , ... , R_k ที่คำนวณจากตัวอย่างที่ลดขนาดต่อเนื่องกันไป คือ n , n-1 , ... , n-k+1 โดยที่ k เป็นจำนวนค่าผิดปกติสูงสุดที่คาดว่าจะมีในชุดข้อมูล 1 ชุด แสดงได้ดังนี้

ข้อมูลเริ่มต้นจะมี x_1 , x_2 , ... , x_n กำหนดให้ $I_0 = \{ x_1 , x_2 , ... , x_n \}$ นั่นคือ I_0 จะมีขนาดตัวอย่างเท่ากับ n จะได้

$$R_1 = \frac{\underset{x_i \in I_0}{\text{Max}} \left| x_i - \bar{x}_1 \right|}{s_1} \qquad \text{โดยที่} \quad \bar{x}_1 = \frac{\sum_{x_i \in I_0} x_i}{n} \qquad \text{และ}$$

$$s_1^{\,2} = \frac{\sum_{x_i \in I_0} \left(x_i - \bar{x}_1 \right)^2}{n-1}$$ จากนั้นตัดค่าสังเกต x_i ที่ให้ค่า $\underset{x_i \in I_0}{\text{Max}} \left| x_i - \bar{x}_1 \right|$

กำหนดให้ $x^{(0)}$ เป็น x_i ที่ $\underset{x_i \in I_0}{Max} \left| x_i - \bar{x}_1 \right|$ และ $I_1 = I_0 - x^{(0)}$ นั่นคือ จากข้อมูลเริ่มต้นจะได้ $x^{(0)}, x^{(1)}, \ldots, x^{(n-1)}$ ซึ่งคือค่าสังเกตที่มีระยะห่างจากค่าเฉลี่ยมากที่สุดในแต่ละ $I_0, I_1, \ldots, I_{n-1}$ และจะได้

$$R_2 = \frac{\underset{x_i \in I_1}{Max} \left| x_i - \bar{x}_2 \right|}{s_2} \qquad \text{โดยที่} \quad \bar{x}_2 = \frac{\sum_{x_i \in I_1} x_i}{n-1} \qquad \text{และ}$$

$$s_2^{\,2} = \frac{\sum_{x_i \in I_1} \left(x_i - \bar{x}_2\right)^2}{n-2} \quad \text{และ } R_3, \ldots, R_k \text{ คำนวนได้ในทำนองเดียวกัน}$$

ค่าวิกฤตของการทดสอบหาได้โดยกำหนด α แล้วหา λ_i , i =1,2,...,k ที่ทำให้

$$\Pr\left\{ \bigcap_{i=L+1}^{k} \left[R_i > \lambda_i\right] \middle| H_L \right\} = \alpha \quad , \qquad L = 0,1,2,\ldots,k-1 \qquad (2.33)$$

วิธีการนี้จะตรวจสอบค่าผิดปกติได้ตั้งแต่ 1 ถึง k ค่า และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างเหมาะสมทั้งภายใต้สมมติฐานหลัก (H_0) คือ ไม่มีค่าผิดปกติในข้อมูล และภายใต้สมมติฐานแย้งของค่าผิดปกติ 1,2,...,k-1 ค่า (H_L , L = 1,2,...,k-1) ตามลำดับ

รูปแบบการพิจารณาค่าผิดปกติของวิธีการ GESD

ถ้าทั้งหมดของ $R_i \le \lambda_i$ แล้วแสดงว่า ไม่มีค่าผิดปกติในข้อมูล

ถ้าบางตัวของ $R_i > \lambda_i$ แล้วกำหนด $C = Max\left\{ i : R_i > \lambda_i \right\}$ จะถือว่า $x^{(0)}, x^{(1)}, \ldots, x^{(C-1)}$ เป็นค่าผิดปกติ เมื่อ $x^{(0)}, x^{(1)}, \ldots, x^{(C-1)}$ เป็นค่าสังเกตที่ให้ค่า $Max\left|x_i - \bar{x}\right|$ ในข้อมูลที่ค่อย ๆ ลดขนาดลงตามลำดับ

จาก (2.33) ต้องการหา λ_i ที่ทำให้ $\Pr\left\{ \underset{i=L+1}{\overset{k}{I}} \left[R_i \le \lambda_i\right] \middle| H_L \right\} = 1-\alpha$

Rosner (1983) ประมาณ $\Pr\left\{ \underset{i=L+1}{\overset{k}{I}} \left[R_i \le \lambda_i\right] \middle| H_L \right\}$ ด้วย $\Pr\left\{ \left[R_{L+1} \le \lambda_{L+1}\right] \middle| H_L \right\}$ เมื่อ L = 0,1,2,...,k-1 ซึ่งการตรวจสอบความถูกต้องของการประมาณนี้ พบว่าไม่เหมาะสมสำหรับกรณีที่ข้อมูลมีขนาดเล็ก (n<25) เพราะระดับนัยสำคัญที่แท้จริงมีค่ามากกว่าระดับนัยสำคัญที่กำหนด อย่างไรก็ตาม เมื่อ n ≥ 25 พบว่าระดับนัยสำคัญที่แท้จริงมีค่าใกล้เคียงหรือ

เท่ากับระดับนัยสำคัญที่กำหนด นั่นคือ ตัวสถิติ GESD ใช้ได้ดีกับชุดข้อมูลที่มีขนาดตัวอย่างมากกว่า 25 ขึ้นไปในการตรวจสอบค่าผิดปกติ

ภายใต้แต่ละสมมติฐาน H_L กำหนด $R_{L+1} = \text{Max}\left\{\left|y_i\right| : i \in I_L\right\}$

โดยที่ $y_i = \dfrac{x_i - \bar{x}^{(L)}}{s^{(L)}}$, $i = 1,2,...,n$, $L = 0,1,...,k-1$

I_L แทนชุดข้อมูลที่ตัดค่าสังเกต x_i ที่ให้ค่า $\underset{x_i \in I_L - 1}{\text{Max}} \left| x_i - \bar{x} \right|$ ออก L ค่า

$\bar{x}^{(L)}$ แทนค่าเฉลี่ยที่ได้จาก I_L

$S^{(L)}$ แทนส่วนเบี่ยงเบนมาตรฐานที่ได้จาก I_L

ค่าวิกฤตของตัวสถิติ GESD (Critical value for GESD statistics)

Grubb (1950) ได้ศึกษาการแจกแจงของ y_i โดยใช้เทคนิคการอินทิเกรตแบบวิธีเชิงตัวเลข (Numerical integration) ซึ่งค่อนข้างยุ่งยาก ดังนั้น เพื่อความสะดวกจึงจะใช้การประมาณของ Quesenberry and David (1961) ทำการประมาณการแจกแจงของ y_i ขั้นแรก ประมาณการแจกแจงของ y_i ซึ่งใช้การแปลงจากการแจกแจงแบบ t ที่ให้ไว้ใน Thomsom (1935) คือ

$$y_i \sim \frac{(n-L-1)t_{n-L-2}}{\left\{\left[n-L-2+t^2_{n-L-2}\right](n-L)\right\}^{\frac{1}{2}}} \quad , \; i \in I_L \qquad (2.34)$$

ขั้นที่สอง ใช้เทคนิค Bonferroni Inequality โดย Quesenberry and David (1961) ประมาณ $\Pr\left(R_{L+1} > \lambda_{L+1}\right)$ ด้วย $(n-L)\Pr\left(y_i > \lambda_{L+1}\right)$ ทำให้ได้

$$\Pr\left(y_i \le \lambda_{L+1}\right) = 1 - \left(\frac{\alpha}{n-L}\right) \qquad (2.35)$$

จาก (2.34) และ (2.35) จะได้ λ_{L+1} สำหรับปัญหาค่าผิดปกติด้านเดียว คือ

$$\lambda_{L+1} = \frac{(n-L-1)t_{d,p}}{\left\{\left[n-L-2+t^2_{d,p}\right](n-L)\right\}^{\frac{1}{2}}} \quad , \quad L=0,1,...,k-1 \qquad (2.36)$$

เมื่อ $p = 1-\left(\frac{\alpha}{n-L}\right)$, $d = n-L-2$ และ $t_{d,p}$ คือ ค่าวิกฤตอันดับที่ p ของการแจกแจงแบบ t ที่มี d เป็นองศาอิสระ

สำหรับปัญหาค่าผิดปกติ 2 ด้าน แทน $\frac{\alpha}{2}$ ลงใน α จะได้

$$\lambda_{L+1} = \frac{(n-L-1)t_{d,p}}{\left\{\left[n-L-2+t^2_{d,p}\right](n-L)\right\}^{\frac{1}{2}}} \quad , \quad L=0,1,...,k-1 \qquad (2.37)$$

โดยที่ $p = 1-\left[\frac{\alpha/2}{n-L}\right]$

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ผู้วิจัยต้องการศึกษาตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติของข้อมูล โดยมีข้อตกลงเบื้องต้นว่า ข้อมูลที่นำมาตรวจสอบค่าผิดปกติส่วนใหญ่มาจากการแจกแจงแบบปกติ ซึ่งการศึกษาครั้งนี้ข้อมูลที่ใช้ทำการศึกษาจะมี 2 ลักษณะ คือ (1) ข้อมูลที่จำลองแบบขึ้น และ (2) ข้อมูลจริงที่คาดว่าน่าจะมีค่าผิดปกติปะปนอยู่ โดยการวิจัยจะใช้การตรวจสอบค่าผิดปกติแบบหลายค่าของ Rosner (1983) คือ ตัวสถิติ Generalized Extreme Studentized Deviate (GESD) ซึ่งเป็นตัวสถิติที่อิงแนวคิดของตัวสถิติแบบดั้งเดิม และตัวสถิติในรูปอย่างง่ายที่พัฒนาขึ้นจากงานวิจัยของ Weiss (1996) ซึ่งอิงแนวคิดของสถิติแบบเบย์ คือ ตัวสถิติคูลแบคก์และตัวสถิติ L_1 ทำการตรวจสอบค่าผิดปกติในข้อมูลที่จำลองแบบและข้อมูลจริงที่นำมาศึกษา โดยผู้วิจัยเสนอวิธีการประมาณตัวสถิติคูลแบคก์และตัวสถิติ L_1 ภายใต้เงื่อนไขต่าง ๆ ดังที่จะเสนอต่อไป

ข้อมูลที่นำมาตรวจสอบค่าผิดปกติ แบ่งได้ 2 ลักษณะ คือ

1. ข้อมูลที่จำลองแบบขึ้น โดยมีขนาดตัวอย่างดังนี้ 10 , 20 , 50 และ 80 จากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน และผสมค่าผิดปกติ 1 ค่าปะปนไปกับข้อมูลที่ได้จากการจำลองโดยค่าผิดปกติดังกล่าวมีขนาดค่อย ๆ เพิ่มมากขึ้นโดยมีขนาดเท่ากับ 3σ , 4σ และ 6σ (โดย $\sigma = 1$) ตามลำดับ

2. ข้อมูลจริงที่คาดว่าน่าจะมีค่าผิดปกติปะปน โดยสมมติว่าข้อมูลส่วนใหญ่มีการแจกแจงแบบปกติซึ่งมีทั้งหมด 3 ชุด ได้แก่

2.1 ข้อมูลของ Freeman (Freeman's Data) Pettit (1992) ได้ศึกษาตัวประกอบเบย์สำหรับตัวแบบผิดปกติโดยใช้วิธีค่าสังเกตจินตภาพ ในการศึกษาตัวประกอบเบย์สำหรับข้อมูลที่มีการแจกแจงแบบปกติ Pettit (1992) ใช้ด้วยข้อมูล 2 ชุด คือ (1) ข้อมูลของ Freeman และ (2) ข้อมูลของ Darwin สำหรับข้อมูลของ Freeman เป็นข้อมูลที่จำลองแบบขึ้นจากการแจกแจงแบบปกติมาตรฐาน โดยข้อมูลแบ่งเป็น 2 เซต เซตแรกประกอบด้วยค่าสังเกตดังนี้ -1.10 -0.28 -0.04 -0.02 0.12 0.71 1.35 1.46 1.74 และ 3.89 และเซตที่สอง แสดงดังตารางที่ 4 ซึ่งเกิดจากการนำข้อมูลในเซตแรก 1 ค่าสังเกต คือ -0.28 ไปบวกกับ -5 โดยถือว่า -5 เป็นค่าผิดปกติที่ปะปนในชุดข้อมูล Pettit (1992) พบว่ามีค่าผิดปกติ 2 ค่าในข้อมูลของ Freeman เซต

ที่สอง คือ –5.28 และ 3.89 และผู้วิจัยจะใช้ข้อมูลของ Freeman เซตที่สองในการตรวจสอบค่าผิดปกติ

2.2 ข้อมูลของ Darwin (Darwin's Data) แสดงดังตารางที่ 5 เป็นข้อมูลผลต่างของความสูงระหว่างพืชที่ผสมในต้น (Self - fertilized) และพืชที่ผสมข้ามต้น (Cross - fertilized) โดยข้อมูลเป็นการวัดความสูงเพื่อเปรียบเทียบ 15 ครั้ง แต่ละครั้งจะเปรียบเทียบการเจริญเติบโตระหว่างพืชที่ผสมในต้นและพืชที่ผสมข้ามต้น (Fisher 1960) ข้อมูลของDarwin ถูกนำมาศึกษาโดย Pettit (1992) เช่นเดียวกับข้อมูลของ Freeman ในการศึกษาของ Pettit (1992) ได้ศึกษาตัวประกอบเบย์สำหรับตัวแบบผิดปกติโดยใช้ค่าสังเกตจินตภาพ ผลการศึกษาพบว่ามีค่าผิดปกติ 2 ค่าในชุดข้อมูล คือ –67 และ -48

2.3 ข้อมูลของ Sacks et al. แสดงดังตารางที่ 6 ข้อมูลของ Sacks et al. เป็นข้อมูลที่ Rosner (1983) นำมาตรวจสอบค่าผิดปกติ โดย Sacks , Omish , Rosner , Mclanahan and Kass ได้ทำการศึกษาในปี 1980 โดยศึกษาจากกลุ่มผู้ที่ทานอาหารมังสะวิรัติจำนวน 54 คน โดยวัดความดันโลหิต และประเมินอาหารที่คนกลุ่มนี้รับประทานโดยใช้แบบสอบถามเกี่ยวกับความถี่ของอาหารที่รับประทานจำนวนมากกว่า 100 ข้อ ตารางแสดงส่วนประกอบมาตรฐานของอาหารถูกนำมาคำนวณคะแนนโภชนาการทั้งหมดตั้งแต่โปรตีน ไขมันและวิตามินต่าง ๆ รวมถึงวิตามิน E ข้อมูลในตารางที่ 6 เป็นข้อมูลคะแนนโภชนาการของวิตามิน E ที่ถูกแปลงเป็น log (log transformation) โดยค่าผิดปกติที่ Rosner (1983) ตรวจพบ คือ 6.01 5.42 และ 5.34

ตารางที่ 4 ข้อมูลของ Freeman

ลำดับที่	ค่าสังเกต	ลำดับที่	ค่าสังเกต
1	**-5.28**	6	0.71
2	-1.10	7	1.35
3	-0.04	8	1.46
4	-0.02	9	1.74
5	0.12	10	**3.89**

ที่มา : Freeman , quoted in L.I. Pettit , "Bayes Factors for Outlier Models Using the Device of Imaginary Observations," Journal of the American Statistical Association 87 (1992) : 542 .

ตารางที่ 5 ข้อมูลของ Darwin

ลำดับที่	ค่าสังเกต	ลำดับที่	ค่าสังเกต
1	-67	9	28
2	-48	10	29
3	6	11	41
4	8	12	49
5	14	13	56
6	16	14	60
7	23	15	75
8	24		

ที่มา : R.A. Fisher , The Design of Experiments , 7th ed. (New York : Hafner Press , 1960) , 37.

ตารางที่ 6 ข้อมูลของ Sacks et al.

ลำดับที่	ค่าสังเกต	ลำดับที่	ค่าสังเกต	ลำดับที่	ค่าสังเกต	ลำดับที่	ค่าสังเกต
1	-0.25	15	1.58	29	2.14	43	2.92
2	0.68	16	1.65	30	2.15	44	2.93
3	0.94	17	1.69	31	2.23	45	3.21
4	1.15	18	1.70	32	2.24	46	3.26
5	1.20	19	1.76	33	2.26	47	3.30
6	1.26	20	1.77	34	2.35	48	3.59
7	1.26	21	1.81	35	2.37	49	3.68
8	1.34	22	1.91	36	2.40	50	4.30
9	1.38	23	1.94	37	2.47	51	4.64
10	1.43	24	1.96	38	2.54	52	**5.34**
11	1.49	25	1.99	39	2.62	53	**5.42**
12	1.49	26	2.06	40	2.64	54	**6.01**
13	1.55	27	2.09	41	2.90		
14	1.56	28	2.10	42	2.92		

ที่มา : Sacks et al. , quoted in B. Rosner , "Precentange Points for a Generalized ESD Many - Outlier Procedure," Technometrics 25(1983) : 171 .

วิธีการและตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติ

การศึกษาจะใช้ 2 วิธีการ ในการตรวจสอบค่าผิดปกติ คือ

1. การตรวจสอบค่าผิดปกติโดยใช้วิธีการทางสถิติแบบดั้งเดิม โดยศึกษาวิธีการตรวจสอบค่าผิดปกติแบบหลายค่า Generalized Extreme Studentized Deviate (GESD) ของ Rosner (1983)

2. วิธีการตรวจสอบค่าผิดปกติแบบทีละค่า โดยใช้การวิเคราะห์ความอ่อนไหวแบบเบย์ ผู้วิจัยเสนอที่จะใช้ตัวสถิติคูลแบคก์และตัวสถิติระยะทาง L_1 ในรูปอย่างง่าย โดยพัฒนาและเสนอวิธีประมาณค่าตัวสถิติทั้ง 2 ดังนี้

2.1 ตัวสถิติวัดการเบี่ยงเบนแบบคูลแบคก์ หรือตัวสถิติแบบคูลแบคก์ (Kullback Statistic) จาก (2.26) จะได้

$$D_1 = \iint -\log_e\left[\frac{p_1\left(\mu,\sigma\middle|x^i\right)}{p\left(\mu,\sigma\middle|\underline{x}\right)}\right]p\left(\mu,\sigma\middle|\underline{x}\right)d\mu\,d\sigma \qquad (3.1)$$

เบื้องต้นจะสมมติว่าทราบความแปรปรวน σ^2 ก่อนจากนั้นจึงประมาณ σ^2 ด้วย S^2 ทั้งนี้เพราะเมื่อหารูปแบบจากการอินทิเกรต 2 ชั้นของ (3.1) ในขั้นแรกโดยโปรแกรมสำเร็จรูป Maple VI ผลที่ได้มีความยุ่งยากและไม่สามารถจัดให้อยู่ในรูปอย่างง่ายได้ ส่งผลต่อวิธีการคำนวณและการนำตัวสถิติแบบคูลแบคก์ไปประยุกต์ใช้ ผู้วิจัยจึงประมาณ D_1 ด้วย K_i เมื่อ

$$K_i = \int_{-\infty}^{\infty} -\log_e\left[\frac{p_1\left(\mu\middle|\sigma,x^i\right)}{p\left(\mu\middle|\sigma,\underline{x}\right)}\right]p\left(\mu\middle|\sigma,\underline{x}\right)d\mu \qquad (3.2)$$

จะได้ $$K_i = \frac{1}{2}+\frac{1}{2}\log_e\left(\frac{n-1}{n}\right)-\frac{n-1}{2\sigma^2}\left(\frac{\sigma^2}{n}+\left(\bar{x}-\bar{x}_i\right)^2\right) \qquad (3.3)$$

ประมาณ σ^2 ด้วย S^2 ใน (3.3) จะได้

$$\tilde{K}_i = \frac{1}{2}+\frac{1}{2}\log_e\left(\frac{n-1}{n}\right)-\frac{n-1}{2s^2}\left(\frac{s^2}{n}+\left(\bar{x}-\bar{x}_i\right)^2\right) \qquad (3.4)$$

2.2 ตัวสถิติระยะทาง L_1 (Weiss 1996) ดังนิยาม (2.32) ที่ปรากฏในบทที่ 2 คือ

$$L_1 = \frac{1}{2}\iint \left| p_1\left(\mu,\sigma \middle| x^i\right) - p\left(\mu,\sigma \middle| \underline{x}\right)\right| d\mu\, d\sigma \qquad (3.5)$$

แทนค่าตัวแบบทั้ง 2 คือ ตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวน ใน (3.5) จะได้

$$L_1 = \frac{1}{2}\int_0^{\infty}\int_{-\infty}^{\infty} \left| \sqrt{\frac{n-1}{2\pi}}\left[\frac{1}{2}\left(\frac{n-2}{2}\right)\right]^{-1}\left(\frac{(n-2)s_i^2}{2}\right)^{\frac{n-2}{2}}\sigma^{-n}\exp\left[-\frac{1}{2}\left\{(n-2)s_i^2+(n-1)(\mu-\bar{x}_i)^2\right\}\right] - \right. \qquad (3.6)$$

$$\left. \sqrt{\frac{n}{2\pi}}\left[\frac{1}{2}\left(\frac{n-1}{2}\right)\right]^{-1}\left(\frac{(n-1)s^2}{2}\right)^{\frac{n-1}{2}}\sigma^{-(n+1)}\exp\left[-\frac{1}{2}\left\{(n-1)s^2+n(\mu-\bar{x})^2\right\}\right]\right| d\mu\, d\sigma$$

จะเห็นว่าต้องอินทิเกรต 2 ชั้นของค่าสัมบูรณ์ที่เกิดจากผลต่างของตัวแบบก่อกวนที่ตัดค่าและตัวแบบสมบูรณ์ วิธีการที่อาจนำมาใช้ในกรณีนี้ คือ เทคนิคการอินทิเกรต 2 ชั้นแบบวิธีเชิงตัวเลข ซึ่งค่อนข้างยุ่งยากและมีอุปสรรคมาก ผู้วิจัยจึงสมมติในเบื้องแรกว่าทราบความแปรปรวนของข้อมูลที่นำมาตรวจสอบค่าผิดปกติ พร้อมสมมติข้อมูลที่นำมาตรวจสอบสุ่มมาจากการแจกแจงแบบปกติและการแจกแจงก่อนของ μ เป็นแบบไม่มีสารสนเทศ ดังนั้นจาก (2.11) และ (2.32) จะได้

$$L_1 = \frac{1}{2}\int_{-\infty}^{\infty}\left| p_1\left(\mu \middle| \sigma, x^i\right) - p\left(\mu \middle| \sigma, \underline{x}\right)\right| d\mu\, d\sigma \qquad (3.7)$$

แทนค่าตัวแบบทั้ง 2 คือ ตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวน ใน (3.7) จะได้

$$L_1 = \frac{1}{2}\int_{-\infty}^{\infty}\left| \sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left[-\frac{n-1}{2\sigma^2}(\mu-\bar{x}_i)^2\right] - \sqrt{\frac{n}{2\pi\sigma^2}}\exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right]\right| d\mu \qquad (3.8)$$

ผู้วิจัยได้ประมาณ (3.8) ใน 2 แนวทางดังนี้

แนวทางแรกประมาณ L_1 ด้วย $\tilde{L}_1$ โดยที่

$$\tilde{L}_1 = \frac{1}{2}\int_{-\infty}^{\infty}\left| p_1\left(\mu \middle| s, x^i\right) - p\left(\mu \middle| s, \underline{x}\right)\right| d\mu \qquad (3.9)$$

หมายความว่า ทั้งตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวน ประมาณ σ^2 ด้วย S^2 เมื่อ S^2 เป็นความแปรปรวนของตัวแบบสมบูรณ์ นั่นคือ ตัวแบบ M_0 : $\mu\left|\sigma^2,\underline{x} \cong \mu\right|s^2,\underline{x} \sim N\left(\bar{x},\frac{s^2}{n}\right)$

และตัวแบบ M_1 : $\mu\left|\sigma^2,x^i \cong \mu\right|s^2,x^i \sim N\left(\bar{x}_i,\frac{s^2}{n-1}\right)$ เมื่อ $\bar{x}_i$ เป็นค่าเฉลี่ยที่ได้จากข้อ

มูลที่ตัดค่าสังเกตตัวที่ i ออก ดังนั้น จาก (3.9) แทนค่าตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวน จะได้

$$\tilde{L}_1 = \frac{1}{2}\int_{-\infty}^{\infty}\left|\sqrt{\frac{n-1}{2\pi s^2}}\exp\left[-\frac{n-1}{2s^2}\left(\mu-\bar{x}_i\right)^2\right]-\sqrt{\frac{n}{2\pi s^2}}\exp\left[-\frac{n}{2s^2}\left(\mu-\bar{x}\right)^2\right]\right|d\mu \quad (3.10)$$

แนวทางที่สองประมาณ L_1 ด้วย $\hat{L}_1$ โดยที่

$$\hat{L}_1 = \int_{-\infty}^{\infty}\left|p_1\left(\mu\middle|s_i,x^i\right)-p\left(\mu\middle|s,\underline{x}\right)\right|d\mu \quad (3.11)$$

หมายความว่า สำหรับตัวแบบสมบูรณ์จะประมาณ σ^2 ด้วย S^2 ส่วนตัวแบบที่ถูกก่อกวนจะประมาณ σ^2 ด้วย S_i^2 เมื่อ S_i^2 คือ ความแปรปรวนของข้อมูลที่ได้จากการตัดค่าสังเกตตัวที่ i ออกจากข้อมูล นั่นคือ ตัวแบบ M_0 : $\mu\left|\sigma^2,\underline{x}\cong\mu\right|s^2,\underline{x}\sim N\left(\bar{x},\frac{s^2}{n}\right)$ และตัวแบบ M_1 : $\mu\left|\sigma^2,x^i\cong\mu\right|s_i^2,x^i\sim N\left(\bar{x}_i,\frac{s_i^2}{n-1}\right)$ ดังนั้น จาก (3.11) แทนค่าตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวน จะได้

$$\hat{L}_1 = \frac{1}{2}\int_{-\infty}^{\infty}\left|\sqrt{\frac{n-1}{2\pi s_i^2}}\exp\left[-\frac{n-1}{2s_i^2}\left(\mu-\bar{x}_i\right)^2\right]-\sqrt{\frac{n}{2\pi s^2}}\exp\left[-\frac{n}{2s^2}\left(\mu-\bar{x}\right)^2\right]\right|d\mu \quad (3.12)$$

ตัวสถิติที่พัฒนาและใช้ในการตรวจสอบค่าผิดปกติของข้อมูลทั่วไป สำหรับการวิจัยครั้งนี้ คือ ตัวสถิติตาม (3.4) , (3.10) และ (3.12)

การพิจารณาค่าผิดปกติ

ในส่วนของการพิจารณาค่าผิดปกติสำหรับตัวสถิติที่อิงแนวคิดแบบเบย์นั้นยังไม่มีรูปแบบที่แน่ชัด ส่วนใหญ่ดูจากค่าสถิติซึ่งเป็นระยะทางของตัวสถิติคูลแบคก์และตัวสถิติ L_1 ที่ได้ ถ้าระยะทางของค่าสังเกตใดมีค่ามากต่างจากระยะทางของค่าสังเกตอื่น ๆ แสดงว่าค่าสังเกตนั้นน่าจะเป็นค่าผิดปกติ เนื่องจากในการศึกษาคุณสมบัติของตัวสถิติ จำเป็นต้องใช้การจำลองแบบโดยคอมพิวเตอร์กับชุดข้อมูลจำนวนมาก จึงจำเป็นต้องมีแนวทางให้คอมพิวเตอร์ตัดสินว่าค่าสังเกตใดเป็นค่าผิดปกติเพื่อจะได้ทราบพฤติกรรมของตัวสถิติต่าง ๆ อย่างคร่าว ๆ ผู้วิจัยจึงจะใช้อัตราส่วนของระยะทางของตัวสถิติที่อิงแนวคิดแบบเบย์ โดยจะเรียกอัตราส่วนดังกล่าวว่าอัตราส่วนระยะทางแบบเบย์ สำหรับการคำนวณอัตราส่วนระยะทางแบบเบย์จะแสดง ดังนี้

สมมติมีค่าสังเกตที่นำมาตรวจสอบค่าผิดปกติทั้งหมด n ค่า คือ a_1 , a_2 , ... , a_n นำค่าสังเกตเหล่านี้ไปคำนวณค่าสถิติที่อิงแนวคิดแบบเบย์ จะได้ค่าสถิติทั้งหมด n ค่า คือ B_1 , B_2 , ... , B_n นำค่าสถิติทั้ง n ค่ามาเรียงลำดับจากน้อยไปมาก จะได้ $B_{[1]}$, $B_{[2]}$, ... , $B_{[n]}$ ตามลำดับ การคำนวณอัตราส่วนระยะทางแบบเบย์จะคำนวณจาก $\frac{B_{[n]} - B_{[n-1]}}{B_{[n-1]} - B_{[n-2]}}$ ถ้าอัตราส่วนระยะทางแบบเบย์มากกว่า 2 คอมพิวเตอร์จะถือว่าค่าสังเกตที่เกิดจากค่าสถิติ $B_{[n]}$ เป็นค่าผิดปกติ นั่นคือ ถ้า $B_{[n]} - B_{[n-1]}$ มีค่ามากกว่า $2(B_{[n-1]} - B_{[n-2]})$ จะถือว่าค่าสังเกตที่ให้ค่าสถิติ $B_{[n]}$ เป็นค่าผิดปกติ และคอมพิวเตอร์จะนับค่าสังเกตนั้นเป็นค่าผิดปกติ เกณฑ์นี้จะใช้เฉพาะกรณีที่ตรวจสอบค่าผิดปกติ 1 ค่า และในส่วนของการพิจารณาค่าผิดปกติที่อิงแนวคิดแบบดั้งเดิมนั้นจะใช้เกณฑ์พิจารณาของ Rosner (1983)

ขั้นตอนการวิจัย จะแบ่งเป็น 2 ส่วน

1. ส่วนของข้อมูลที่จำลองแบบขึ้น โดยจะมีขั้นตอนดังนี้

1.1 สร้างข้อมูลที่มีการแจกแจงแบบปกติมาตรฐานโดยการจำลองแบบทั้งหมด 4 ขนาดตัวอย่าง คือ 10 , 20 , 50 และ 80

1.2 กำหนดค่าผิดปกติ 3 ขนาด คือ 3σ , 4σ และ 6σ (σ = 1) และปะปนค่าผิดปกติลงในข้อมูลที่จำลองแบบขึ้น

1.3 ใช้ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม คือ ตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}'_1$ ตรวจสอบค่าผิดปกติที่ปะปนลงไป

1.4 ทำซ้ำขั้นตอนที่ 1.1 ถึง 1.3 ทั้งหมด 2,000 ครั้ง และตรวจสอบว่าใน 2,000 ครั้ง นั้นสามารถตรวจสอบค่าผิดปกติได้ทั้งหมดกี่ครั้งและคิดเป็นร้อยละเท่าไร

หมายเหตุ ในการจำลองแบบอาจเกิดกรณีที่มีค่าผิดปกติ ไม่ได้เกิดจากการกำหนดของผู้วิจัยปะปนมาจากการจำลองแบบ ดังนั้น ในการวิจัยจะตรวจสอบด้วยว่าค่าผิดปกติที่ตรวจพบนั้น เป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัยกี่ค่า

2. ส่วนของข้อมูลจริงที่นำมาตรวจสอบ มีขั้นตอนดังนี้

2.1 ใช้ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม คือ ตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}'_1$ ตรวจสอบค่าผิดปกติในชุดข้อมูลจริง โดยจะตรวจสอบว่ามีค่าผิดปกติกี่ค่า ค่าสังเกตใดบ้างที่เป็นค่าผิดปกติ

2.2 เปรียบเทียบตัวสถิติทั้งหมดว่า ให้ผลการตรวจสอบนำไปสู่ข้อสรุปเดียวกันหรือไม่

บทที่ 4
ผลการวิจัย

การตรวจสอบค่าผิดปกติในงานวิจัยนี้ ในส่วนของข้อมูลจะมี 2 ลักษณะ คือ (1) ข้อมูลที่จำลองแบบขึ้นจากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน และ (2) ข้อมูลจริงที่นำมาศึกษา ในส่วนของตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติเป็นตัวสถิติที่อิงแนวคิดแบบดั้งเดิม คือ ตัวสถิติ Generalized Extreme Studentized Deviate (GESD) และตัวสถิติที่อิงแนวคิดแบบเบย์ คือ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1'$ ในส่วนของข้อมูลที่จำลองแบบ ค่าผิดปกติที่ปะปนลงไปมี 3 ขนาด คือ ขนาดเล็ก (3σ) ขนาดกลาง (4σ) และขนาดใหญ่ (6σ) โดยที่ $\sigma = 1$ สำหรับการนำเสนอผลการวิจัยผู้วิจัยจะนำเสนอตามลักษณะของข้อมูลดังนี้

1. ข้อมูลที่ได้จากการจำลองแบบมี 4 ขนาดตัวอย่าง คือ 10 , 20 , 50 และ 80
2. ข้อมูลจริงที่นำมาตรวจสอบค่าผิดปกติ คือ ข้อมูลของ Freeman ข้อมูลของ Darwin และข้อมูลของ Sacks et al.

ข้อมูลจากการจำลองแบบ

สำหรับข้อมูลที่ได้จากการจำลองแบบ บางครั้งค่าผิดปกติที่ตรวจพบอาจไม่ได้เกิดจากการกำหนดของผู้วิจัย ดังนั้น ในการนำเสนอจะแยกค่าผิดปกติที่ตรวจพบเป็น 2 แบบ คือ แบบ A ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติทั้งที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย และแบบ B ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย

ขนาดตัวอย่างเท่ากับ 10

ตารางที่ 7 จำนวนครั้งที่ตรวจพบค่าผิดปกติในตัวอย่างขนาด 10 จากการทำซ้ำ 2000 ครั้ง ของตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05 , $\tilde{K}$, $\tilde{L}_1$ และ l_1 โดยการปะปนค่าผิดปกติทีละค่า ที่ค่าผิดปกติขนาด 3, 4 และ 6

		ตัวสถิติ							
ค่าผิดปกติ		GESD		$\tilde{K}$		$\tilde{L}_1$		l_1	
ขนาด	แบบ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
3	A	676	33.80	1327	66.35	1143	57.13	1308	65.40
	B	672	33.60	1318	65.90	1134	56.70	1299	64.95
4	A	1460	73.00	1718	85.90	1522	76.10	1700	85.00
	B	1460	73.00	1718	85.90	1522	76.10	1700	85.00
6	A	1994	99.70	1969	98.45	1839	91.95	1966	98.30
	B	1994	99.70	1969	98.45	1839	91.95	1966	98.30

ร้อยละ : ค่าร้อยละจำนวนครั้งที่ตรวจพบค่าผิดปกติ จากการทำซ้ำ 2000 ครั้ง

จากตารางที่ 7 สามารถแบ่งได้ 2 กรณี ดังนี้

1. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย (แบบ A)

1.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ l_1 สามารถตรวจพบค่าผิดปกติได้ 676 , 1327 , 1143 และ 1308 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD , $\tilde{L}_1$, l_1 และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 33.80 , 57.13 , 65.40 และ 66.35 ตามลำดับ

1.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1460 , 1718 , 1522 และ 1700 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD, $\tilde{L}_1$, $\tilde{L}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 73.00 , 76.10 , 85.00 และ 85.90 ตามลำดับ

1.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1994 , 1969 , 1839 และ 1966 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 91.95 , 98.30 , 98.45 และ 99.70 ตามลำดับ

2. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย (แบบ B)

1.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 672 , 1318 , 1134 และ 1299 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD, $\tilde{L}_1$, $\tilde{L}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 33.60 , 56.70 , 64.95 และ 65.90 ตามลำดับ

1.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1460 , 1718 , 1522 และ 1700 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD, $\tilde{L}_1$, $\tilde{L}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 73.00 , 76.10 , 85.00 และ 85.90 ตามลำดับ

1.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1994 , 1969 , 1839 และ 1966 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 91.95 , 98.30 , 98.45 และ 99.70 ตามลำดับ

ขนาดตัวอย่างเท่ากับ 20

ตารางที่ 8 จำนวนครั้งที่ตรวจพบค่าผิดปกติในตัวอย่างขนาด 20 จากการทำซ้ำ 2000 ครั้ง ของตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05 , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ โดยการปะปนค่าผิดปกติทีละค่า ที่ค่าผิดปกติขนาด 3, 4 และ 6

		ตัวสถิติ							
ค่าผิดปกติ		GESD		$\tilde{K}$		$\tilde{L}_1$		$\acute{L}_1$	
ขนาด	แบบ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
3	A	698	34.90	1266	63.30	1113	55.65	1203	60.15
	B	674	33.70	1242	62.10	1096	54.80	1181	59.05
4	A	1773	88.65	1756	87.80	1572	78.60	1688	84.40
	B	1770	88.50	1756	87.80	1572	78.60	1688	84.40
6	A	2000	100.0	1986	99.30	1908	95.40	1964	98.20
	B	2000	100.0	1986	99.30	1908	95.40	1964	98.20

ร้อยละ : ค่าร้อยละจำนวนครั้งที่ตรวจพบค่าผิดปกติ จากการทำซ้ำ 2000 ครั้ง

จากตารางที่ 8 สามารถแบ่งได้ 2 กรณี ดังนี้

1. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย (แบบ A)

1.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 698 , 1266 , 1113 และ 1203 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD, $\tilde{L}_1$, $\acute{L}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 34.90 , 55.65 , 60.15 และ 63.30 ตามลำดับ

1.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1773 , 1756 , 1572 และ 1688 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 78.60 , 84.40 , 87.80 และ 88.65 ตามลำดับ

1.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 2000 , 1986 , 1908 และ 1964 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 95.40 , 98.20 , 99.30 และ 100.00 ตามลำดับ

2. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย (แบบ B)

2.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 674 , 1242 , 1096 และ 1181 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD, $\tilde{L}_1$, $\tilde{L}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 33.70 , 54.80 , 59.35 และ 62.10 ตามลำดับ

2.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1770 , 1756 , 1572 และ 1688 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 78.60 , 84.40 , 87.80 และ 88.50 ตามลำดับ

2.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 2000 , 1986 , 1908 และ 1964 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 95.40 , 98.20 , 99.30 และ 100.00 ตามลำดับ

ขนาดตัวอย่างเท่ากับ 50

ตารางที่ 9 จำนวนครั้งที่ตรวจพบค่าผิดปกติในตัวอย่างขนาด 50 จากการทำซ้ำ 2000 ครั้ง ของตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05 , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ โดยการปะปนค่าผิดปกติทีละค่า ที่ค่าผิดปกติขนาด 3, 4 และ 6

ค่าผิดปกติ		ตัวสถิติ							
		GESD		$\tilde{K}$		$\tilde{L}_1$		$\tilde{l}_1$	
ขนาด	แบบ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
3	A	522	26.10	1136	56.80	1041	52.05	1067	53.35
	B	427	21.35	1068	53.40	987	49.35	1009	50.45
4	A	1994	99.70	1728	86.40	1572	78.60	1628	81.40
	B	1934	96.70	1728	86.40	1572	78.60	1628	81.40
6	A	2000	100.0	1987	99.35	1934	96.70	1965	98.25
	B	2000	100.0	1987	99.35	1934	96.70	1965	98.25

ร้อยละ : ค่าร้อยละจำนวนครั้งที่ตรวจพบค่าผิดปกติ จากการทำซ้ำ 2000 ครั้ง

จากตารางที่ 9 สามารถแบ่งได้ 2 กรณี ดังนี้

1. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย (แบบ A)

1.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ สามารถตรวจพบค่าผิดปกติได้ 522 , 1136 , 1041 และ 1067 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD , $\tilde{L}_1$, $\tilde{l}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 26.10 , 52.05 , 53.35 และ 56.80 ตามลำดับ

1.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ สามารถตรวจพบค่าผิดปกติได้ 1994 , 1728 , 1572 และ 1628 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{l}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 78.60 , 81.40 , 86.40 และ 99.70 ตามลำดับ

1.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ สามารถตรวจพบค่าผิดปกติได้ 2000 , 1987 , 1934 และ 1965 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{l}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 96.70 , 98.25 , 99.35 และ 100.00 ตามลำดับ

2. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย (แบบ B)

2.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ สามารถตรวจพบค่าผิดปกติได้ 427 , 1068 , 987 และ 1009 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD , $\tilde{L}_1$, $\tilde{l}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 21.35 , 49.35 , 50.45 และ 53.40 ตามลำดับ

2.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ สามารถตรวจพบค่าผิดปกติได้ 1934 , 1728 , 1572 และ 1628 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{l}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 78.60 , 81.40 , 86.40 และ 96.70 ตามลำดับ

2.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ สามารถตรวจพบค่าผิดปกติได้ 2000 , 1987 , 1934 และ 1965 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\tilde{l}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 96.70 , 98.25 , 99.35 และ 100.00 ตามลำดับ

ขนาดตัวอย่างเท่ากับ 80

ตารางที่ 10 จำนวนครั้งที่ตรวจพบค่าผิดปกติในตัวอย่างขนาด 80 จากการทำซ้ำ 2000 ครั้ง ของตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05 , $\tilde{K}$, $\tilde{L}_1$ และ l_1 โดยการปะปนค่าผิดปกติทีละค่า ที่ค่าผิดปกติขนาด 3, 4 และ 6

ค่าผิดปกติ		ตัวสถิติ							
		GESD		$\tilde{K}$		$\tilde{L}_1$		l_1	
ขนาด	แบบ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
3	A	314	15.70	996	49.80	922	46.10	937	46.85
	B	214	10.70	915	45.75	850	42.50	864	43.20
4	A	1978	98.90	1700	85.00	1587	79.35	1618	80.90
	B	1968	98.40	1700	85.00	1587	79.35	1618	80.90
6	A	2000	100.0	1992	99.60	1957	97.85	1976	98.80
	B	2000	100.0	1992	99.60	1957	97.85	1976	98.80

ร้อยละ : ค่าร้อยละจำนวนครั้งที่ตรวจพบค่าผิดปกติ จากการทำซ้ำ 2000 ครั้ง

จากตารางที่ 10 สามารถแบ่งได้ 2 กรณี ดังนี้

1. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย (แบบ A)

1.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ l_1 สามารถตรวจพบค่าผิดปกติได้ 314 , 996 , 922 และ 937 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD , $\tilde{L}_1$, l_1 และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 15.70 , 46.10 , 46.85 และ 49.80 ตามลำดับ

1.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1978 , 1700 , 1587 และ 1618 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\acute{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 79.35 , 80.90 , 85.00 และ 98.90 ตามลำดับ

1.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 2000 , 1992 , 1957 และ 1976 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\acute{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 97.85 , 98.80 , 99.60 และ 100.00 ตามลำดับ

2. กรณีที่ค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย (แบบ B)

2.1 ค่าผิดปกติขนาดเล็ก (เท่ากับ 3)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 214 , 915 , 850 และ 864 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ GESD , $\tilde{L}_1$, $\acute{L}_1$ และ $\tilde{K}$ ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 10.70 , 42.50 , 43.20 และ 45.75 ตามลำดับ

2.2 ค่าผิดปกติขนาดกลาง (เท่ากับ 4)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 1968 , 1700 , 1587 และ 1618 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, $\acute{L}_1$, $\tilde{K}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 79.35 , 80.90 , 85.00 และ 98.40 ตามลำดับ

2.3 ค่าผิดปกติขนาดใหญ่ (เท่ากับ 6)

ตัวสถิติ GESD , $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้ 2000 , 1992 , 1957 และ 1976 ครั้ง ตามลำดับ จากการทำซ้ำ 2000 ครั้ง เมื่อเรียงลำดับตัวสถิติที่

ตรวจพบค่าผิดปกติจากน้อยไปมากจะได้ $\tilde{L}_1$, L_1 , $\tilde{\kappa}$ และ GESD ซึ่งมีร้อยละในการตรวจพบค่าผิดปกติคิดเป็น 97.85 , 98.80 , 99.60 และ 100.00 ตามลำดับ

ข้อมูลจริง

สำหรับข้อมูลจริง ผู้วิจัยสนใจว่า ตัวสถิติที่นำมาตรวจสอบค่าผิดปกติ จะให้ผลลัพธ์ที่นำไปสู่ข้อสรุปเดียวกันหรือไม่ ดังนั้น ในการนำเสนอจะนำเสนอตามตัวสถิติที่อิง 2 แนวคิด คือ ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม คือ ตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{\kappa}$, $\tilde{L}_1$ และ L_1 ตามลำดับ และเป็นที่ทราบกันแล้วว่าค่าผิดปกติที่ตรวจพบในชุดข้อมูลจะมีไม่เกินครึ่งหนึ่งของข้อมูลทั้งหมด ดังนั้น จะแสดงข้อมูลค่าสังเกตและตัวสถิติที่คำนวณจากแต่ละชุดข้อมูลเพียงครึ่งหนึ่งของทั้งหมด โดยตัวสถิติ GESD จะพิจารณาตามเกณฑ์ Rosner (1983) ส่วนตัวสถิติที่อิงแนวคิดแบบเบย์จะพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต ระยะทางของค่าสังเกตใดห่างจากค่าสังเกตอื่นมากจะถือว่าค่าสังเกตนั้นเป็นค่าผิดปกติ

ข้อมูลของ Freeman

1. ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม (ตัวสถิติ GESD)

ตารางที่ 11 ค่าสถิติและค่าวิกฤตที่คำนวณจากข้อมูล Freeman โดยตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05

ครั้งที่	ขนาดตัวอย่าง	ค่าสถิติ GESD	ค่าวิกฤต	ค่าสังเกต
1	10	2.33872	2.28995	-5.28
2	9	2.07891	2.21501	3.89
3	8	1.69074	2.12665	-1.10
4	7	1.29070	2.01997	1.74
5	6	1.26238	1.88715	1.46

จากตารางที่ 11 จะพบว่าข้อมูลของ Freeman มีค่าผิดปกติ 1 ค่า คือ –5.28 ทั้งนี้เพราะค่าสถิติ GESD ตัวที่ 1 เท่ากับ 2.33872 ซึ่งมากกว่า ค่าวิกฤต คือ 2.28995

2. ตัวสถิติที่อิงแนวคิดแบบเบย์

ตารางที่ 12 ระยะทางของตัวสถิติที่อิงแนวคิดแบบเบย์ที่คำนวณจากข้อมูล Freeman โดยตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$

$\tilde{K}$		$\tilde{L}_1$		$\acute{L}_1$	
ระยะทาง	ค่าสังเกต	ระยะทาง	ค่าสังเกต	ระยะทาง	ค่าสังเกต
0.30655	-5.28	0.31146	-5.28	0.41667	-5.28
0.13043	3.89	0.20563	3.89	0.21677	3.89
0.02352	1.74	0.08595	1.74	0.08889	1.74
0.02146	-1.10	0.08185	-1.10	0.08550	-1.10
0.01628	1.46	0.07050	1.46	0.07678	1.46

จากตารางที่ 12 จะพิจารณาตามตัวสถิติดังนี้

2.1 ตัวสถิติ $\tilde{K}$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Freeman น่าจะมีค่าผิดปกติ 2 ค่า คือ –5.28 และ 3.89 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต –5.28 และ 3.89 ซึ่งเท่ากับ 0.30655 และ 0.13043 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

2.2 ตัวสถิติ $\tilde{L}_1$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Freeman น่าจะมีค่าผิดปกติ 2 ค่า คือ –5.28 และ 3.89 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต –5.28 และ 3.89 ซึ่งเท่ากับ 0.31146 และ 0.20563 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

2.3 ตัวสถิติ $\acute{L}_1$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Freeman น่าจะมีค่าผิดปกติ 2 ค่า คือ –5.28 และ 3.89 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่า

สังเกต –5.28 และ 3.89 ซึ่งเท่ากับ 0.41667 และ 0.21677 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

ข้อมูลของ Darwin

1. ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม (ตัวสถิติ GESD)

ตารางที่ 13 ค่าสถิติและค่าวิกฤตที่คำนวณจากข้อมูล Darwin โดยตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05

ครั้งที่	ขนาดตัวอย่าง	ค่าสถิติ GESD	ค่าวิกฤต	ค่าสังเกต
1	15	2.32971	2.54831	-67
2	14	2.51140	2.50732	-48
3	13	1.94980	2.46204	75
4	12	1.67276	2.41156	60
5	11	1.80087	2.35473	56
6	10	1.83379	2.28995	49
7	9	1.79425	2.21501	41
8	8	1.41795	2.12665	6

จากตารางที่ 13 จะพบว่าข้อมูลของ Darwin มีค่าผิดปกติ 2 ค่า คือ –67 และ –48 ทั้งนี้เพราะค่าสถิติ GESD ตัวที่ 2 เท่ากับ 2.51140 ซึ่งมากกว่า ค่าวิกฤตตัวที่ 2 คือ 2.50732

2. ตัวสถิติที่อิงแนวคิดแบบเบย์

ตารางที่ 14 ระยะทางของตัวสถิติที่อิงแนวคิดแบบเบย์ที่คำนวณจากข้อมูล Darwin โดยตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1'$

$\tilde{K}$		$\tilde{L}_1$		$\tilde{L}_1'$	
ระยะทาง	ค่าสังเกต	ระยะทาง	ค่าสังเกต	ระยะทาง	ค่าสังเกต
0.19500	-67	0.24881	-67	0.28612	-67
0.12029	-48	0.19649	-48	0.20886	-48
0.07444	75	0.15495	75	0.15820	75
0.03942	60	0.11269	60	0.11277	60
0.03199	56	0.10140	56	0.10141	56
0.02091	49	0.08164	49	0.08223	49
0.01126	41	0.05916	41	0.06163	41
0.00675	6	0.04495	6	0.05017	6

จากตารางที่ 14 จะพิจารณาตามตัวสถิติดังนี้

2.1 ตัวสถิติ $\tilde{K}$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Darwin น่าจะมีค่าผิดปกติ 2 ค่า คือ –67 และ -48 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต –67 และ -48 ซึ่งเท่ากับ 0.195000 และ 0.12029 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

2.2 ตัวสถิติ $\tilde{L}_1$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Darwin น่าจะมีค่าผิดปกติ 2 ค่า คือ –67 และ -48 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต –67 และ -48 ซึ่งเท่ากับ 0.24881 และ 0.19649 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

2.3 ตัวสถิติ L_1

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Darwin น่าจะมีค่าผิดปกติ 2 ค่า คือ –67 และ -48 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต –67 และ -48 ซึ่งเท่ากับ 0.28612 และ 0.20886 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

ข้อมูลของ Sacks et al.

1. ตัวสถิติที่อิงแนวคิดแบบดั้งเดิม (ตัวสถิติ GESD)

ตารางที่ 15 ค่าสถิติและค่าวิกฤตที่คำนวณจากข้อมูล Sacks et al. โดยตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05

ครั้งที่	ขนาดตัวอย่าง	ค่าสถิติ GESD	ค่าวิกฤต	ค่าสังเกต
1	54	3.11891	3.15879	6.01
2	53	2.94297	3.15143	5.42
3	52	3.17942	3.14389	5.34
4	51	2.81018	3.13617	4.64
5	50	2.81558	3.12825	-0.25
6	49	2.84817	3.12013	4.30
7	48	2.27933	3.11180	3.68
8	47	2.31037	3.10324	3.59
9	46	2.10158	3.09446	0.68
10	45	2.06718	3.08543	3.30
11	44	2.13373	3.07614	3.26
12	43	2.19555	3.06658	3.21
13	42	1.91447	3.05672	0.94

14	41	1.85469	3.04657	2.93
15	40	1.94558	3.03610	2.92
16	39	2.07575	3.02528	2.92

ตารางที่ 15 (ต่อ)

ครั้งที่	ขนาดตัวอย่าง	ค่าสถิติ GESD	ค่าวิกฤต	ค่าสังเกต
17	38	2.19219	3.01411	2.90
18	37	1.77799	3.00255	2.64
19	36	1.84077	2.99059	2.62
20	35	1.76585	2.97819	2.54
21	34	1.72908	2.96532	1.15
22	33	1.70804	2.95195	1.20
23	32	1.70681	2.93805	2.47
24	31	1.65602	2.92357	1.26
25	30	1.76804	2.90847	1.26
26	29	1.65883	2.89270	1.34
27	28	1.65176	2.87621	1.38

จากตารางที่ 15 จะพบว่าข้อมูลของ Sacks et al. มีค่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 ทั้งนี้เพราะค่าสถิติ GESD ตัวที่ 3 เท่ากับ 3.17942 ซึ่งมากกว่า ค่าวิกฤตตัวที่ 3 คือ 3.14389

2. ตัวสถิติที่อิงแนวคิดแบบเบย์

ตารางที่ 16 ระยะทางของตัวสถิติที่อิงแนวคิดแบบเบย์ที่คำนวณจากข้อมูล Sacks et al. โดยตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ ℓ_1

$\tilde{K}$		$\tilde{L}_1$		ℓ_1	
ระยะทาง	**ค่าสังเกต**	**ระยะทาง**	**ค่าสังเกต**	**ระยะทาง**	**ค่าสังเกต**
0.09186	6.01	0.17040	6.01	0.18128	6.01
0.06485	5.42	0.14349	5.42	0.14920	5.42

0.06155	5.34	0.13982	5.34	0.14500	5.34
0.04465	-0.25	0.11924	-0.25	0.12198	-0.25
0.03635	4.64	0.10766	4.64	0.10942	4.64

ตารางที่ 16 (ต่อ)

$\tilde{K}$		$\tilde{L}_1$		$\tilde{l}_1$	
ระยะทาง	**ค่าสังเกต**	**ระยะทาง**	**ค่าสังเกต**	**ระยะทาง**	**ค่าสังเกต**
0.02650	4.30	0.09197	4.30	0.09280	4.30
0.01824	0.68	0.07632	0.68	0.07659	0.68
0.01294	0.94	0.06429	0.94	0.06434	0.94
0.01254	3.68	0.06414	3.68	0.06334	3.68
0.01095	3.59	0.05913	3.59	0.05914	3.59
0.00933	1.15	0.05456	1.15	0.05456	1.15
0.00856	1.20	0.05225	1.20	0.05225	1.20
0.00767	1.26	0.04947	1.26	0.04948	1.26
0.00767	1.26	0.04947	1.26	0.04948	1.26
0.00657	1.34	0.04576	1.34	0.04581	1.34
0.00655	3.30	0.04570	3.30	0.04575	3.30
0.00605	1.38	0.04391	1.38	0.04397	1.38
0.00604	3.26	0.04384	3.26	0.04390	3.26
0.00544	1.43	0.04160	1.43	0.04169	1.43
0.00542	3.21	0.04153	3.21	0.04162	3.21
0.00474	1.49	0.03882	1.49	0.03895	1.49
0.00474	1.49	0.03882	1.49	0.03895	1.49
0.00409	1.55	0.03605	1.55	0.03623	1.55
0.00399	1.56	0.03558	1.56	0.03578	1.56
0.00379	1.58	0.03475	1.58	0.03487	1.58
0.00312	1.65	0.03142	1.65	0.03172	1.65
0.00277	1.69	0.02958	1.69	0.02993	1.69

จากตารางที่ 16 จะพิจารณาตามตัวสถิติดังนี้

2.1 ตัวสถิติ $\tilde{K}$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Sacks et al. น่าจะมีค่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต 6.01 , 5.42 และ 5.34 ซึ่งเท่ากับ 0.09186 , 0.06485 และ 0.06155 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

2.2 ตัวสถิติ $\tilde{L}_1$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Sacks et al. น่าจะมีค่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต 6.01 , 5.42 และ 5.34 ซึ่งเท่ากับ 0.17040 , 0.14349 และ 0.13982 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

2.3 ตัวสถิติ $L_1^{\prime}$

จากระยะทางของตัวสถิติ สามารถพิจารณาได้ว่าข้อมูลของ Sacks et al. น่าจะมีค่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 โดยพิจารณาจากระยะทางของตัวสถิติที่เกิดจากค่าสังเกต 6.01 , 5.42 และ 5.34 ซึ่งเท่ากับ 0.18128 , 0.14920 และ 0.14500 ตามลำดับ ซึ่งห่างจากระยะทางที่เกิดจากค่าสังเกตอื่น ๆ

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

สรุปผลการวิจัย

การวิจัยนี้ได้เสนอตัวสถิติที่ใช้ตรวจสอบค่าผิดปกติขึ้นมาใหม่ทั้งหมด 3 ตัวคือ $\tilde{K}$, $\tilde{L}_1$ และ l_1 ซึ่งเป็นตัวสถิติที่อิงแนวคิดแบบเบย์ โดยพัฒนาจากงานวิจัยของ Weiss (1996) ทั้ง 3 ตัวสถิติใช้การแจกแจงก่อนแบบไม่มีสารสนเทศตามแนวทางของ Jeffrey ซึ่งผู้วิจัยเห็นว่าสามารถนำผลการตรวจสอบค่าผิดปกติไปเปรียบเทียบกับตัวสถิติที่อิงแนวคิดแบบดั้งเดิม คือ ตัวสถิติ Generalized Extreme Studentized Deviate (GESD) ซึ่ง Rosner (1983) ได้เสนอไว้ การตรวจค่าผิดปกติครั้งนี้ กระทำภายใต้ข้อสมมติว่าข้อมูลตัวอย่างที่นำมาตรวจสอบค่าผิดปกตินั้น มาจากประชากรที่มีการแจกแจงแบบปกติ โดยข้อมูลจะมี 2 ลักษณะ คือ (1) ข้อมูลที่จำลองแบบขึ้นที่มีขนาดตัวอย่าง 4 ขนาด คือ 10 , 20 , 50 และ 80 และ (2) ข้อมูลจริงที่คาดว่าน่าจะมีค่าผิดปกติปะปน คือ ข้อมูลของ Freeman ข้อมูลของ Darwin และข้อมูลของ Sacks et al.

ในส่วนข้อมูลที่จำลองแบบขึ้น จะปะปนค่าผิดปกติ 3 ขนาด คือ ค่าผิดปกติขนาดเล็ก ค่าผิดปกติขนาดกลาง และค่าผิดปกติขนาดใหญ่ การตรวจสอบค่าผิดปกติจากตัวสถิติที่อิงแนวคิดแบบเบย์และตัวสถิติที่อิงแนวคิดดั้งเดิม การตรวจสอบจะทำโดยแยกเป็นค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย (แบบ A) และค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย (แบบ B) ซึ่งสามารถสรุปผลการวิจัยได้ดังนี้

ข้อมูลจากการจำลองแบบ

1. ขนาดตัวอย่าง

1.1 ขนาดตัวอย่างเท่ากับ 10

กรณีที่ค่าผิดปกติมีขนาดเล็ก (3σ) และขนาดกลาง (4σ) สามารถสรุปได้ว่าตัวสถิติ $\tilde{K}$ มีจำนวนครั้งในการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ l_1 , $\tilde{L}_1$ และ GESD ตามลำดับ กรณีที่ค่าผิดปกติมีขนาดเล็ก ตัวสถิติ GESD สามารถตรวจพบค่าผิดปกติได้น้อยมาก เนื่องจากมีจำนวนครั้งในการตรวจพบค่าผิดปกติ เมื่อเทียบกับตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ l_1 น้อยที่สุด ซึ่งเห็นได้ชัดจากร้อยละของการตรวจพบค่าผิดปกติจากตารางที่ 7 กรณีค่าผิดปกติ

มีขนาดกลางก็สรุปผลได้เช่นเดียวกับกรณีค่าผิดปกติมีขนาดเล็ก แต่จำนวนครั้งที่ตรวจพบค่าผิดปกติสำหรับกรณีค่าผิดปกติขนาดกลางไม่ต่างกันมากเหมือนกรณีค่าผิดปกติมีขนาดเล็ก

กรณีที่ค่าผิดปกติมีขนาดใหญ่ (6σ) สามารถสรุปได้ว่าตัวสถิติ GESD มีจำนวนครั้งในการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ $\tilde{K}$, L_1' และ $\tilde{L}_1$ ตามลำดับ ซึ่งมีข้อน่าสังเกต คือ ในกรณีที่ค่าผิดปกติมีขนาดเล็กและขนาดกลาง ตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ L_1' สามารถตรวจสอบค่าผิดปกติได้มากกว่าตัวสถิติ GESD แต่กรณีที่ค่าผิดปกติขนาดใหญ่ ซึ่งน่าจะตรวจพบค่าผิดปกติได้ดี กลับตรวจสอบได้ไม่ดีเท่าที่ควร

1.2 ขนาดตัวอย่างเท่ากับ 20 , 50 และ 80

กรณีที่ค่าผิดปกติมีขนาดเล็ก (3σ) สามารถสรุปได้ว่าตัวสถิติ $\tilde{K}$ มีจำนวนครั้งในการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ L_1', $\tilde{L}_1$ และ GESD ตามลำดับ ผลการวิจัยสรุปได้เช่นเดียวกับในกรณีขนาดตัวอย่างเท่ากับ 10 นั่นคือ มีความแตกต่างอย่างเห็นได้ชัดในการตรวจพบค่าผิดปกติของตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ โดยตัวสถิติที่อิงแนวคิดแบบเบย์ทั้ง 3 ตัว สามารถตรวจพบค่าผิดปกติได้มากกว่าตัวสถิติ GESD

กรณีที่ค่าผิดปกติมีขนาดกลาง (4σ) และขนาดใหญ่ (6σ) สามารถสรุปได้ว่าตัวสถิติ GESD มีจำนวนครั้งในการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ $\tilde{K}$, L_1' และ $\tilde{L}_1$ ตามลำดับ จะเห็นว่าผลการวิจัยมีข้อน่าสังเกตเช่นเดียวกับกรณีที่ค่าผิดปกติมีขนาดใหญ่ในขนาดตัวอย่างเท่ากับ 10

จาก 1.1 และ 1.2 สำหรับตัวสถิติที่อิงแนวคิดแบบเบย์ จะสรุปว่าตัวสถิติ $\tilde{K}$ มีความสามารถในการตรวจพบค่าผิดปกติมากที่สุด ถัดมาคือ ตัวสถิติ L_1' และ $\tilde{L}_1$ ตามลำดับ

2. ขนาดของค่าผิดปกติ

ในภาพรวมจะพบว่าขนาดของค่าผิดปกติ มีผลต่อการตรวจพบค่าผิดปกติในข้อมูลตัวอย่างที่จำลองแบบขึ้น ทั้งนี้เพราะในตัวอย่างชุดเดียวกันแต่ขนาดของค่าผิดปกติต่างกัน คือ ค่าผิดปกติขนาดเล็ก ค่าผิดปกติขนาดกลางและค่าผิดปกติขนาดใหญ่ จะพบว่าค่าผิดปกติขนาดเล็กจะสามารถตรวจพบได้น้อยกว่าค่าผิดปกติขนาดกลางและค่าผิดปกติขนาดใหญ่ และจะตรวจพบค่าผิดปกติขนาดกลางได้น้อยกว่าค่าผิดปกติขนาดใหญ่ ซึ่งพบในทุกขนาดตัวอย่างที่จำลองแบบขึ้น

2.1 ค่าผิดปกติขนาดเล็ก

ตัวสถิติ GESD มีแนวโน้มที่จะตรวจพบค่าผิดปกติลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น สังเกตจากร้อยละในการตรวจพบแบบ B ร้อยละในการตรวจพบเท่ากับ 33.60 , 33.70 , 21.35 และ 10.70 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ส่วนตัวสถิติที่อิงแนวคิดแบบเบย์ ผลการตรวจพบค่าผิดปกติจะให้ผลลัพธ์เช่นเดียวกับตัวสถิติ GESD นั่นคือ ตัวสถิติ $\tilde{K}$ มีร้อยละในการตรวจพบค่าผิดปกติแบบ B เท่ากับ 65.90 , 62.10 , 53.40 และ 45.75 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ตัวสถิติ $\tilde{L}_1$ มีร้อยละในการตรวจพบค่าผิดปกติแบบ B เท่ากับ 56.70 , 54.80 , 49.35 และ 42.50 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ และตัวสถิติ l_1 มีร้อยละในการตรวจพบค่าผิดปกติแบบ B เท่ากับ 64.95 , 59.05 , 50.45 และ 43.20 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ นั่นคือ ถ้าขนาดตัวอย่างเพิ่มขึ้นแนวโน้มที่จะตรวจพบค่าผิดปกติลดลง ทั้งในตัวสถิติ GESD และตัวสถิติทั้ง 3 ที่อิงแนวคิดแบบเบย์ สำหรับการตรวจพบค่าผิดปกติแบบ A จะให้ผลลัพธ์ในทำนองเดียวกันกับการตรวจพบค่าผิดปกติแบบ B

2.2 ค่าผิดปกติขนาดกลาง

ตัวสถิติ GESD มีแนวโน้มที่จะตรวจพบค่าผิดปกติเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น สังเกตจากร้อยละในการตรวจพบแบบ B ร้อยละในการตรวจพบเท่ากับ 73.00 , 88.50 , 96.70 และ 98.40 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ร้อยละในการตรวจพบของตัวสถิติ $\tilde{K}$ ในการตรวจพบแบบ B คือ 85.90 , 87.80 , 86.40 และ 85.00 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ร้อยละในการตรวจพบของตัวสถิติ $\tilde{L}_1$ ในการตรวจพบแบบ B คือ 76.10 , 78.60 , 78.60 และ 79.30 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ และร้อยละในการตรวจพบของตัวสถิติ l_1 ในการตรวจพบแบบ B คือ 85.00 , 84.00 , 81.40 และ 80.90 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ดังนั้นในตัวสถิติที่อิงแนวคิดแบบเบย์จะสรุปได้ว่าตัวสถิติ $\tilde{K}$ และ $\tilde{L}_1$ มีแนวโน้มในการตรวจพบค่าผิดปกติค่อนข้างคงที่ในการตรวจสอบค่าผิดปกติทุกขนาดตัวอย่าง ส่วนตัวสถิติ l_1 นั้น เมื่อตัวอย่างมีขนาดเพิ่มขึ้นร้อยละในการตรวจพบมีแนวโน้มลดลง สำหรับการตรวจพบค่าผิดปกติแบบ A จะให้ผลลัพธ์ในทำนองเดียวกันกับการตรวจพบค่าผิดปกติแบบ B

2.3 ค่าผิดปกติขนาดใหญ่

ตัวสถิติ GESD มีความคงที่ค่อนข้างมากในการตรวจพบค่าผิดปกติเมื่อขนาดตัวอย่างเปลี่ยนไป โดยร้อยละในการตรวจพบแบบ B เท่ากับ 99.70 , 100.00 , 100.00 และ 100.00 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ในส่วนตัวสถิติที่อิงแนวคิดแบบเบย์ ก็ค่อนข้างจะให้ผลการตรวจพบค่าผิดปกติที่คงที่ เช่นเดียวกับตัวสถิติ GESD นั่นคือ ตัวสถิติ $\tilde{K}$ มีร้อยละในการตรวจพบค่าผิดปกติแบบ B เท่ากับ 98.45 , 99.30 , 99.35 และ 99.60 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ ตัวสถิติ $\tilde{L}_1$ มีร้อยละในการตรวจพบค่าผิดปกติแบบ B เท่ากับ 91.95 , 95.40 , 96.70 และ 97.85 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ และตัวสถิติ L_1 มีร้อยละในการตรวจพบค่าผิดปกติแบบ B เท่ากับ 98.30 , 98.20 , 98.25 และ 98.80 เมื่อขนาดตัวอย่างเท่ากับ 10 , 20 , 50 และ 80 ตามลำดับ สำหรับการตรวจพบค่าผิดปกติแบบ A จะให้ผลลัพธ์เหมือนกันกับการตรวจพบค่าผิดปกติแบบ B

3. รูปแบบในการตรวจสอบค่าผิดปกติ

ดังที่ได้กล่าวไว้ก่อนหน้านี้ ว่าการวิจัยนี้จะทำการตรวจสอบด้วยว่าค่าผิดปกติที่ตรวจพบนั้นเกิดจากการกำหนดของผู้วิจัยหรือไม่ โดยให้แบบ A เป็นการตรวจพบค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย และแบบ B เป็นการตรวจพบค่าผิดปกติที่เกิดจากการกำหนดของผู้วิจัย ผลการวิจัยสรุปได้ว่าที่ค่าผิดปกติมีขนาดเล็ก (3σ) ทุกขนาดตัวอย่างที่พิจารณา การตรวจพบค่าผิดปกติแบบ A สูงกว่าแบบ B ในทุกตัวสถิติ นั่นคือ สำหรับค่าผิดปกติขนาดเล็ก มีการตรวจพบค่าผิดปกติที่ไม่ได้เกิดจากการกำหนดของผู้วิจัยค่อนข้างสูง นั่นคือ กรณีค่าผิดปกติขนาดเล็ก อาจมีการตรวจพบค่าผิดปกติที่ไม่ได้เกิดจากการกำหนดของผู้วิจัยค่อนข้างสูง ซึ่งพบในทุกตัวสถิติ

ในกรณีที่ค่าผิดปกติมีขนาดกลาง (4σ) ตัวสถิติที่อิงแนวคิดแบบเบย์ตรวจพบค่าผิดปกติแบบ A และแบบ B เท่ากัน นั่นคือ ตัวสถิติที่อิงแนวคิดแบบเบย์ จะไม่มีการตรวจพบค่าผิดปกติที่ไม่ได้เกิดจากการกำหนดของผู้วิจัย กล่าวคือ ค่าผิดปกติที่ตรวจพบทุกค่าเกิดจากการกำหนดของผู้วิจัย ส่วนตัวสถิติ GESD เมื่อขนาดตัวอย่างเท่ากับ 10 การตรวจพบค่าผิดปกติแบบ A และแบบ B เท่ากัน แต่ในขนาดตัวอย่าง 20 , 50 และ 80 ตัวสถิติ GESD ตรวจพบค่าผิดปกติแบบ A มากกว่าแบบ B โดยตรวจพบค่าผิดปกติแบบ A ร้อยละ 88.65 , 99.70 และ 98.90 เมื่อ

ขนาดตัวอย่างเท่ากับ 20 , 50 และ 80 ตามลำดับ แต่ตรวจพบค่าผิดปกติแบบ B ร้อยละ 88.50 , 96.70 และ 98.40 เมื่อขนาดตัวอย่างเท่ากับ 20 , 50 และ 80 ตามลำดับ นั่นคือ กรณีค่าผิดปกติขนาดกลางและขนาดตัวอย่างเท่ากับ 10 ค่าผิดปกติที่ตรวจพบจากตัวสถิติ GESD ทุกค่าเกิดจากการกำหนดของผู้วิจัย แต่ในขนาดตัวอย่างขนาด 20 , 50 และ 80 ค่าผิดปกติที่ตรวจพบบางครั้งไม่ได้เกิดจากการกำหนดของผู้วิจัย

ในกรณีค่าผิดปกติมีขนาดใหญ่ (6σ) ไม่มีความแตกต่างในผลการตรวจพบค่าผิดปกติแบบ A และแบบ B ของทุกตัวสถิติที่นำมาตรวจสอบ นั่นคือ ค่าผิดปกติที่ตรวจพบทุกค่าเกิดจากการกำหนดของผู้วิจัย ในทุกตัวสถิติ

ข้อมูลจริง

1. ข้อมูลของ Freeman

ตัวสถิติ GESD ตรวจพบค่าผิดปกติ คือ –5.28 เป็นค่าผิดปกติเพียงค่าเดียวที่ระดับนัยสำคัญ 0.05 ส่วนตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ ตรวจพบค่าผิดปกติ 2 ค่า คือ –5.28 และ 3.89 ดังนั้น สามารถสรุปได้ว่าในข้อมูล Freeman ตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ให้ข้อสรุปที่แตกต่างกันในการตรวจพบค่าผิดปกติ

2. ข้อมูลของ Darwin

ตัวสถิติ GESD ตรวจพบค่าผิดปกติ 2 ค่า คือ –67 และ –48 ที่ระดับนัยสำคัญ 0.05 และตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ พบค่าผิดปกติ 2 ค่า คือ –67 และ –48 เช่นกัน ดังนั้น สามารถสรุปได้ว่าในข้อมูลของ Darwin ตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ให้ข้อสรุปในการตรวจพบค่าผิดปกติไปในทำนองเดียวกัน

3. ข้อมูลของ Sacks et al.

ตัวสถิติ GESD ตรวจพบค่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 ที่ระดับนัยสำคัญ 0.05 และตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ พบค่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 เช่นกัน ดังนั้น สามารถสรุปได้ว่าในข้อมูล Sacks et al. ตัวสถิติ GESD และตัวสถิติที่อิงแนวคิดแบบเบย์ให้ข้อสรุปในการตรวจพบค่าผิดปกติไปในทำนองเดียวกัน

อภิปรายผล

1. จากการสรุปผลการวิจัย จะพบข้อน่าสังเกต เกิดขึ้นในข้อมูลที่จำลองแบบ คือ กรณีตัวอย่างขนาดเท่ากับ 10 และกำหนดค่าผิดปกติขนาดใหญ่ และกรณีตัวอย่างขนาด 20 , 50 และ 80 และกำหนดค่าผิดปกติขนาดกลางและขนาดใหญ่ ตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{\kappa}$, $\tilde{L}_1$ และ ℓ_1 มีความสามารถในการตรวจพบค่าผิดปกติน้อยกว่าตัวสถิติ GESD ทั้งที่ในกรณีกำหนดค่าผิดปกติขนาดเล็ก ตัวสถิติที่อิงแนวคิดแบบเบย์มีความสามารถในการตรวจพบค่าผิดปกติมากกว่าตัวสถิติ GESD อย่างเห็นได้ชัดจากร้อยละในการตรวจพบที่ห่างถึง 2-4 เท่า เพื่อไขข้อสงสัยนี้ ผู้วิจัยจึงพิมพ์ผลของการจำลองข้อมูลและค่าสถิติต่าง ๆ (ดูภาคผนวก ข) และพบว่า

1.1 ตัวอย่างชุดเดียวกัน กรณีกำหนดค่าผิดปกติขนาดเล็ก และตัวสถิติที่อิงแนวคิดแบบเบย์สามารถตรวจพบค่าผิดปกติ หากเปลี่ยนจากค่าผิดปกติขนาดเล็กเป็นขนาดกลางหรือขนาดใหญ่ ตัวสถิติที่อิงแนวคิดแบบเบย์ก็ยังสามารถตรวจพบค่าผิดปกติได้เช่นกัน กรณีนี้ค่าผิดปกติที่ตรวจพบเกิดจากการกำหนดของผู้วิจัย

1.2 ตัวอย่างชุดเดียวกัน กรณีกำหนดค่าผิดปกติขนาดเล็ก และตัวสถิติที่อิงแนวคิดแบบเบย์สามารถตรวจพบค่าผิดปกติ แต่ค่าผิดปกติที่ตรวจพบไม่ได้เกิดจากการกำหนดของผู้วิจัย จะได้ว่าเมื่อกำหนดค่าผิดปกติเป็นขนาดกลางหรือขนาดใหญ่ ตัวสถิติที่อิงแนวคิดแบบเบย์ บางครั้งอาจจะตรวจไม่พบค่าผิดปกติ ทั้งนี้เกิดจากอัตราส่วนระยะทางแบบเบย์ของค่าสังเกตที่คำนวณออกมานั้น น้อยกว่า 2 เพื่อให้เห็นภาพชัดเจน จะสมมติว่าในข้อมูลของ Freeman ในตารางที่ 12 มีค่าผิดปกติเพียง 1 ค่า

พิจารณาข้อมูลของ Freeman จะได้ระยะทางของตัวสถิติ $\tilde{\kappa}$ คือ 0.30655 0.13043 0.02352 0.02146 และ 0.01628 ซึ่งเกิดจากค่าสังเกต –5.28 3.89 1.74 -1.10 และ 1.46 ตามลำดับ เมื่อพิจารณาแล้วจะได้ค่าผิดปกติ 2 ค่า คือ –5.28 และ 3.89 ต่อไปคำนวณอัตราส่วนระยะทางแบบเบย์ของค่าสังเกต –5.28 จาก (0.30655 – 0.13043)/(0.13043 – 0.02352) = 1.647367 จะเห็นว่าอัตราส่วนระยะทางแบบเบย์ของค่าสังเกต –5.28 เท่ากับ 1.647367 น้อยกว่า 2 ดังนั้น ถ้าบอกว่าในข้อมูลของ Freeman มีค่าผิดปกติ 1 ค่า จะสรุปจากอัตราส่วนระยะทางแบบเบย์ได้ว่า –5.28 ไม่เป็นค่าผิดปกติและข้อมูลของ Freeman ไม่มีค่าผิดปกติ แต่ถ้าคำนวณอัตราส่วนระยะแบบเบย์ของค่าสังเกต 3.89 จาก (0.13043 – 0.02352)/

(0.02352 – 0.02146) = 51.89806 จะเห็นว่าอัตราส่วนระยะทางแบบเบย์ของค่าสังเกต 3.89 เท่ากับ 51.89806 มากกว่า 2 น่าจะสรุปได้ว่ามีค่าผิดปกติ 2 ค่าในข้อมูล Freeman

จาก 1.2 ประเด็นปัญหาที่เกิดกับตัวสถิติที่อิงแนวคิดแบบเบย์ ในการจำลองแบบข้อมูล คือ เกณฑ์ในการตัดสินว่าค่าสังเกตใดเป็นค่าผิดปกติยังไม่สมบูรณ์นัก กล่าวคือ กรณีที่ค่าผิดปกติมีมากกว่า 1 ค่า ซึ่งเป็นผลที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัยในการจำลองแบบ (ดูผลลัพธ์จากการจำลองแบบในภาคผนวก ข) เมื่อพิจารณาจากระยะทางของตัวสถิติแล้วน่าจะมีค่าผิดปกติมากกว่า 1 ค่าทั้งที่ผู้วิจัยกำหนดค่าผิดปกติเพียง 1 ค่า ทำให้ตรวจไม่พบค่าผิดปกติขนาดใหญ่ในข้อมูลบางชุด ไม่ได้เกิดจากตัวสถิติที่ประมาณขึ้น แต่เกิดจากผลของเกณฑ์ในการพิจารณาค่าผิดปกติยังไม่สมบูรณ์เมื่อมีค่าผิดปกติมากกว่า 1 ค่า หรือที่เรียกว่าการแอบแฝงของค่าผิดปกติในชุดข้อมูล (Masking effect) ในทางปฏิบัติ เมื่อนำตัวสถิติที่อิงแนวคิดแบบเบย์ไปใช้ตรวจสอบค่าผิดปกติกับข้อมูลจริงปัญหาดังกล่าวจะไม่เกิดขึ้น เพราะไม่จำเป็นต้องใช้เกณฑ์อัตราส่วนระยะทางแบบเบย์ในการพิจารณาค่าผิดปกติ โดยสามารถพิจารณาจากค่าสถิติที่คำนวณได้ทันที เกณฑ์อัตราส่วนระยะทางแบบเบย์เป็นเพียงแนวทางให้คอมพิวเตอร์ตัดสินค่าผิดปกติในการจำลองแบบเท่านั้น

2. ชุดข้อมูลจริงที่นำมาศึกษาค่าผิดปกติมี 3 ชุด คือ ข้อมูลของ Freeman ข้อมูลของ Darwin และข้อมูลของ Sacks et al.

Pettit (1992) ตรวจสอบค่าผิดปกติในข้อมูลของ Freeman พบค่าผิดปกติ 2 ค่า คือ -5.28 และ 3.89 เมื่อนำข้อมูลของ Freeman ไปตรวจสอบค่าผิดปกติด้วยตัวสถิติที่อิงแนวคิดแบบเบย์ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ พบว่า -5.28 และ 3.89 เป็นค่าผิดปกติเช่นกัน แต่เมื่อนำตัวสถิติ GESD ไปตรวจสอบค่าผิดปกติที่ระดับนัยสำคัญ 0.05 พบว่า -5.28 เป็นค่าผิดปกติเพียงค่าเดียว ที่เป็นเช่นนี้ อาจเนื่องมาจากข้อมูลของ Freeman มีขนาดตัวอย่างเล็ก ซึ่ง Rosner (1983) ได้อภิปรายไว้ว่าตัวสถิติ GESD เหมาะสำหรับขนาดตัวอย่างที่ใหญ่พอสมควร จึงจะทำให้ระดับนัยสำคัญที่แท้จริงเท่ากับระดับนัยสำคัญที่กำหนด นอกจากนี้ Pettit (1992) ยังได้ตรวจสอบค่าผิดปกติในข้อมูลของ Darwin พบค่าผิดปกติ 2 ค่า คือ -67 และ -48 เมื่อนำตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ พบว่า -67 และ -48 เป็นค่าผิดปกติ และเมื่อนำตัวสถิติ GESD ไปตรวจสอบค่าผิดปกติที่ระดับนัยสำคัญ 0.05 พบว่า -67 และ -48 เป็นค่าผิดปกติเช่นกัน

สำหรับข้อมูลของ Sacks et al. เป็นข้อมูลที่ Rosner (1983) นำมาตรวจสอบค่าผิดปกติด้วยตัวสถิติ GESD ที่ระดับนัยสำคัญ 0.05 พบว่าผิดปกติ 3 ค่า คือ 6.01 , 5.42 และ 5.34 เมื่อนำตัวสถิติที่ได้จากการวิจัยในครั้งนี้ คือ ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{l}_1$ ไปตรวจสอบค่าผิด

ปกติ พบว่า 6.01 , 5.42 และ 5.34 เป็นค่าผิดปกติเช่นเดียวกัน จะเห็นได้ว่าข้อมูลของ Darwin และข้อมูลของ Sacks et al. มีขนาดใหญ่ว่าข้อมูลของ Freeman ซึ่งมีผลต่อการตรวจพบค่าผิดปกติสำหรับตัวสถิติ GESD

3. สำหรับตัวสถิติที่อิงแนวคิดแบบเบย์ ในข้อมูลที่จำลองแบบขึ้นจะพบว่าตัวสถิติ $\tilde{K}$ มีจำนวนครั้งมากที่สุดในการตรวจพบค่าผิดปกติ ถัดมาคือ ตัวสถิติ $\acute{L}_1$ และ $\tilde{L}_1$ ตามลำดับ และเมื่อพิจารณาชุดข้อมูลจริงที่นำมาตรวจสอบค่าผิดปกติ พบว่าตัวสถิติที่อิงแนวคิดแบบเบย์ทั้ง 3 ตัว สถิติให้ผลลัพธ์ไปในทำนองเดียวกัน ทั้ง 3 ชุดข้อมูล เมื่อเปรียบเทียบตัวสถิติ $\tilde{L}_1$ และ $\acute{L}_1$ ซึ่งตัวสถิติทั้งสองได้จากการประมาณตัวสถิติระยะทาง L_1 ที่มีรูปแบบการประมาณที่ต่างกัน ในส่วนของความแปรปรวนของข้อมูลตัวอย่างที่ใช้ประมาณความแปรปรวนของประชากรในตัวแบบที่ถูกก่อกวน กล่าวคือ $\acute{L}_1$ จะตัดค่าสังเกตตัวที่ i ออกก่อน แล้วคำนวณความแปรปรวนจากตัวอย่างเพื่อประมาณความแปรปรวนของประชากรในตัวแบบที่ถูกก่อกวน ดังนั้น ตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวนในตัวสถิติ $\acute{L}_1$ จะพิจารณาความแปรปรวนที่แตกต่างกัน ขณะที่ตัวสถิติ $\tilde{L}_1$ ใช้ค่าสังเกตครบทุกค่าในการคำนวณความแปรปรวนจากข้อมูลตัวอย่าง เพื่อประมาณความแปรปรวนของประชากร ในตัวแบบที่ถูกก่อกวน ดังนั้น ตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวนในตัวสถิติ $\tilde{L}_1$ มีความแปรปรวนที่เท่ากัน สำหรับชุดข้อมูลจริงตัวสถิติ $\acute{L}_1$ ค่อนข้างพิจารณาค่าผิดปกติได้ง่ายกว่าตัวสถิติ $\tilde{L}_1$ เพราะระยะทางของตัวสถิติมีความต่างอย่างชัดเจน ซึ่งสอดรับกับผลที่ได้จากการจำลองแบบข้อมูล ซึ่งพบว่าตัวสถิติ $\acute{L}_1$ สามารถตรวจพบค่าผิดปกติได้มากกว่าตัวสถิติ $\tilde{L}_1$ แสดงว่าการพิจารณาความแปรปรวนในชุดข้อมูลทำให้ระยะทางของตัวแบบสมบูรณ์และตัวแบบที่ถูกก่อกวนโดยการตัดค่าเพิ่มขึ้นส่งผลต่อการตรวจพบค่าผิดปกติ ผลของการวิจัยนี้ทำให้สามารถประมาณค่าของตัวสถิติ L_1 ได้ง่ายขึ้น โดยใช้ตัวสถิติ $\tilde{L}_1$ และ $\acute{L}_1$ แทน

ข้อเสนอแนะ

1. ตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ $\acute{L}_1$ เป็นการตรวจสอบค่าผิดปกติแบบตัดค่าสังเกตทีละค่า อาจเกิดการแอบแฝงค่าผิดปกติในชุดข้อมูลได้ ซึ่งอาจเป็นไปได้ที่จะตรวจสอบค่าผิดปกติแบบตัดค่าสังเกตทีละหลายค่า สมมติเป็น k ค่า โดยอาจจะหาความน่าจะเป็นในการเกิดค่าผิดปกติก่อนแล้วดูว่าถ้าตัดค่าสังเกตเหล่านั้นออกพร้อมกัน ผลที่ได้เป็นอย่างไร ซึ่งข้อมูลที่ใช้ในการจำลองแบบมีค่าผิดปกติเพียงค่าเดียว ทำให้ไม่เห็นลักษณะดังกล่าว และในข้อมูลจริงมีค่าสังเกตที่เป็นค่าผิดปกติหลายค่า แต่ไม่ได้วิเคราะห์ลักษณะตำแหน่งของค่าผิดปกติว่ามีผลต่อการวิเคราะห์อย่างไร จึงอาจเป็นอีกข้อเสนอแนะให้วิจัยต่อไปได้

2. การพิจารณาค่าผิดปกติในตัวสถิติ $\tilde{K}$, $\tilde{L}_1$ และ L_1' ในกรณีข้อมูลจำลองแบบเป็นการพิจารณาจากอัตราส่วนระยะทางแบบเบย์ที่มีค่ามากกว่า 2 ซึ่งจำเป็นต้องใช้เป็นแนวทางคร่าว ๆ ให้คอมพิวเตอร์ตรวจสอบค่าผิดปกติ ซึ่งอาจศึกษาต่อไปว่ามีเกณฑ์อื่นที่เหมาะกว่าอัตราส่วนระยะทางแบบเบย์หรือไม่

3. ในการพัฒนาตัวสถิติ ผู้วิจัยประมาณความแปรปรวนของประชากร (σ^2) ด้วยความแปรปรวนของข้อมูลตัวอย่าง (S^2) ค่าประมาณความแปรปรวนที่ได้ไม่มีความแกร่ง (Robust) ต่อค่าผิดปกติ ซึ่งอาจส่งผลต่อตัวสถิติ แนวทางในการวิจัยต่อไป อาจลองใช้ตัวสถิติที่มีความแกร่งต่อค่าผิดปกติมากกว่า S^2 เพื่อประมาณ σ^2 เช่น Trimmed variance ส่วนเบี่ยงเบนเฉลี่ย (Mean of the Absolute Deviation : MAD) เมื่อ MAD = $\frac{\sum|x_i - \bar{x}|}{n}$ หรือตัวสถิติ $\frac{\sum|x_i - \text{Median}|}{n}$ เป็นต้น

บรรณานุกรม

Box , G.E.P., and G.C. Tiao. Bayesian Inference in Statistical Analysis. California : Addison-Wesley Publisting Company , 1973.

Brant , R. "Comparing Classical and Resistant Outlier Rules." Journal of the American Statistical Association 85 (1990) : 1083-1090.

Carey , V.J. et al. "Resistant and Test-Based Outlier Rejection : Effects on Gaussian One- and Two- Sample Inference." Technometrics 39 (1997) : 320-330.

Carlin , B.P., and N.G. Polson. "An Expected Utility Approach to Influence Diagnostics." Journal of the American Statistical Association 86 (1991) : 1013-1021.

Davies , L., and U. Gather. "The Identification of Multiple Outliers." Journal of the American Statistical Association 88 (1993) : 782-792.

Dixon , W.J. "Processing Data for Outliers." Biometrics 9 (1953) : 482-493.

Ferguson , T.S. "On the Rejection of Outliers." Proceedings of the Fourth Berkeley Symposium 1 (1961) : 253-287.

Fisher , R.A. The Design of Experiments. 7th ed. New York : Hafner Press , 1960.

_____ . Statistical Methods Experimental Design and Scientific Inference. New York : Oxford University Press , 1990.

Geisser , S. "Discussion on Sampling and Bayes' Inference In Scientific Modelling And Robustness." Journal of the Royal Statistical Society 143 (1980) : 416-417.

_____ . "Predictive Discordancy tests for Exponential Observations." The Canadian Journal of Statistics 17 (1989) : 19-26.

_____ . Predictive Inference : an Introduction. London : Chapman and Hall , 1993.

Grubbs , F.E. "Sample Criteria for Testing Outlying Observation." Annals of Mathematical Statistics 29 (1950) : 27-58.

Gustafson , P. "On Measuring Sensitivity to Parametric Model Misspecification." Journal of the Royal Statistical Society 63 (2001) : 82-94.

Hawkins , D. "Letter to the Editor." Technometrics 20 (1978) : 218.

Johnson , W., and S. Geisser. "A Predictive View of the Detection and Characterization of Influential Observation in Regression Analysis." Journal of the American Statistical Association 78 (1983) : 137-144.

Kass , R.E. "Bayes Factors in Practices." Statistician 42 (1993) : 551-560.

Kass , R.E., and A.E. Raftery. "Bayes Factors." Journal of the American Statistical Association 90 (1990) : 773-795.

Kass , R.E. , L. Tierney , and J.B. Kadane. "Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis." Biometrika 76 (1989) : 663-674.

McCulloch , R.E. "Local Model Influence." Journal of the American Statistical Association 84 (1989) : 473-478.

McMillan , R.G. "Tests for One or Two Outliers in Normal Samples with Unknown Variance." Technometrics 49 (1971) : 87-100.

Murphy , R.B. "Procedures for Detecting Outlying in Two-samples." Ph.D Dissertation , Prince University , 1951.

O'Connor , J.J., and E.F. Robertson. William Sealey Gosset [Online] . Accessed 3 September 2002. Available from http://www-history.mcs.st-andrews.ac.uk/history/Mathematicians/Gosset.html

Paulson ,E. "An Optimum Solution to the k-Sample Slippage Problem for Normal Distribution." Annals of Mathematical Statistics 23 (1952) : 610-616.

Pettit , L.I. "Bayes Methods for Outliers in Exponential Sample." Journal of the Royal Statistical Society 50 (1988) : 371-380.

_____ . "The Conditional Predictive Ordinate for the Normal Distribution." Journal of the Royal Statistical Society 52 (1990) : 175-184.

Pettit , L.I. "Bayes Factors for Outlier Models Using the Device of Imaginary Observations." Journal of the American Statistical Association 87 (1992) : 541-545.

Pettit , L.I., and A.F.M. Smith. Bayesian Statistic 2. Amsterdam : North - Holland, 1985.

Prescott , P. "Critical Values for a Sequential Test for Many Outliers." Applied Statistics 28 (1979) : 36-39.

Press , J.S. Bayesian Statistics : Principles , Models and Application. New York : John Wiley&Sons , 1989.

Quesanberry , C.P., and H.A. David. "Some Test for Outliers." Biometrika 48 (1961) : 379-390.

Rosner , B. "On the Detection of Many Outliers." Technometrics 17 (1975) : 221-227.

_____ . "Precentange Points for a Generalized ESD Many-Outlier Procedure." Technometrics 25(1983) : 165-172.

Stigler , S.M. The History of Statistics . Cambridge : President and Fellows of Harvard College , 1986.

Thompson , W.R. "On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation." Annals of Mathematical Statistics 6 (1935) : 214-219.

Tietjen , G.L., and R.H. Moore. "Some Grubbs-type Statistics for the Detection of Several Outliers." Technometrics 55 (1972) : 583-298.

Tukey , J.W. Exploratory Data Analysis. Reading , MA : Addison - Wesley , 1977.

Verdinelli , I., and L. Wasserman. "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ration." Journal of the American Statistical Association 90 (1995) : 614-618.

Weiss , N.A. Elementary Statistics. 2nd ed. Massachusetts : Addison - Wesley Publishing Company , 1993

Weiss , R. "An Apporoach to Bayesian Sensitivity Analysis." Journal of the Royal Statistical Society 58 (1996) : 739-750.

Weiss , R.E., and R.D. Cook. "A Graphical Case Statistic for Assessing Posterior Influence." Biometrika 79 (1992) : 51-55.

Young , K.D.S., and L.I. Pettit. "Measuring Discordancy between Prior and Data." Journal of the Royal Statistical Society 58 (1996) : 679-689.

Zhu , H.T., and S.Y. Lee. "Local Influence for Incomplete-data Model." Journal of the Royal Statistical Society 63 (2001) : 111-126.

ภาคผนวก ก

การพิสูจน์การแจกแจงภายหลังของ θ เมื่อกำหนด y

ที่มาของการแจกแจงภายหลัง (Posterior Distribution) ของ θ เมื่อกำหนด y

กำหนดการแจกแจงก่อน (Prior Distribution) ของ θ คือ $N(\theta_0, \sigma_0^2)$ นั่นคือ

$$p(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right] \quad , \quad -\infty < \theta < \infty \qquad (ก1.1)$$

และ $\theta \sim N(\theta, \sigma^2)$ โดยที่ ฟังก์ชันภาวะน่าจะเป็น (Likelihood function) ของ θ คือ เขียนในรูปแปรผัน (proportional) กับการแจกแจงแบบปกติ ดังนี้

$$L(\theta|y) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta-y}{\sigma}\right)\right] \qquad (ก1.2)$$

ดังนั้น การแจกแจงภายหลัง) ของ θ เมื่อกำหนด y คือ

$$p(\theta|y) = \frac{p(\theta)L(\theta|y)}{\int_{-\infty}^{\infty} p(\theta)L(\theta|y)\,d\theta}$$

$$= \frac{f(\theta|y)}{\int_{-\infty}^{\infty} f(\theta|y)\,d\theta} \quad , \quad -\infty < \theta < \infty \qquad (ก1.3)$$

เมื่อ $f(\theta|y) = \exp\left\{-\frac{1}{2}\left[\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2 + \left(\frac{y-\theta}{\sigma}\right)^2\right]\right\}$ (ก1.4)

พิจารณา

$$A(z-a)^2 + B(z-b)^2 = (A+B)(z-c)^2 + \frac{AB}{A+B}(a-b)^2 \qquad (ก.1.5)$$

เมื่อ $c = \frac{1}{A+B}(Aa+Bb)$

ดังนั้น จะได้

$$\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2+\left(\frac{y-\theta}{\sigma}\right)^2 \quad = \quad (\sigma_0^{-2}+\sigma^{-2})(\theta-\bar{\theta})^2+d$$

เมื่อ $\bar{\theta} \quad = \quad \frac{1}{\sigma_0^{-2}+\sigma^{-2}}(\sigma_0^{-2}\theta_0+\sigma^{-2}y)$

และ d เป็นค่าคงที่ที่ไม่เกี่ยวข้องกับ θ ดังนั้น

$$f(\theta|y) \quad = \quad \exp(d)\exp\left\{-\frac{1}{2}(\sigma_0^{-2}+\sigma^{-2})(\theta-\bar{\theta})^2\right\} \qquad (ก1.6)$$

และ

$$\int_{-\infty}^{\infty} f(\theta|y)\,d\theta \quad = \quad \exp(d)\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}(\sigma_0^{-2}+\sigma^{-2})(\theta-\bar{\theta})^2\right\}d\theta$$

$$= \quad \sqrt{2\pi}\,(\sigma_0^{-2}+\sigma^{-2})^{-1/2}\exp(-d/2) \qquad (ก1.7)$$

ดังนั้น จะได้

$$p(\theta|y) \quad = \quad \frac{(\sigma_0^{-2}+\sigma^{-2})^{1/2}}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(\sigma_0^{-2}+\sigma^{-2})(\theta-\bar{\theta})^2\right\} \quad ,$$

$$-\infty \quad < \theta < \quad \infty \qquad (ก1.8)$$

นั่นคือ (ก1.8) เป็นการแจกแจงแบบปกติ $N(\bar{\theta}, (\sigma_0^{-2}+\sigma^{-2})^{-1})$

ภาคผนวก ข
ข้อมูลบางส่วนที่ได้จากการจำลองแบบ

ขนาดตัวอย่างเท่ากับ 20 ค่าผิดปกติขนาดเล็ก (3σ)

ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	$\tilde{L}_1'$
1	0.1017	0.0007	0.1017	0.0125	0.1017	0.0255
2	0.221	0.0008	0.221	0.0135	0.221	0.026
3	0.0181	0.0008	0.0181	0.0141	0.0181	0.0262
4	-0.0122	0.0009	-0.0122	0.0151	-0.0122	0.0266
5	-0.0818	0.0013	-0.0818	0.018	-0.0818	0.028
6	0.415	0.0018	0.415	0.0219	0.415	0.0301
7	0.4256	0.0019	0.4256	0.0225	0.4256	0.0304
8	-0.1892	0.0021	-0.1892	0.0239	-0.1892	0.0312
9	0.5466	0.0031	0.5466	0.0299	0.5466	0.035
10	-0.3422	0.0038	-0.3422	0.0335	-0.3422	0.0377
11	0.6455	0.0044	0.6455	0.0363	0.6455	0.0398
12	0.848	0.0079	0.848	0.0496	0.848	0.0512
13	-0.7067	0.0105	-0.7067	0.0575	-0.7067	0.0585
14	-0.839	0.0138	-0.839	0.0663	-0.839	0.0669
15	-0.8829	0.0151	-0.8829	0.0692	-0.8829	0.0697
16	-0.9552	0.0172	-0.9552	0.0741	-0.9552	0.0743
17	1.6721	0.034	1.6721	0.1045	1.6721	0.1047
18	2.4004	0.073	2.4004	0.153	2.4004	0.1574
19	3	0.1162	3	0.1926	3	0.2058
20	-3.6755	0.2039	-3.6755	0.2534	-3.6755	0.2964

ขนาดตัวอย่างเท่ากับ 20 ค่าผิดปกติขนาดกลาง (4σ)

ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	L_1
1	0.221	0.0007	0.221	0.0126	0.221	0.0256
2	0.1017	0.0007	0.1017	0.0131	0.1017	0.0258
3	0.0181	0.001	0.0181	0.0153	0.0181	0.0267
4	-0.0122	0.0011	-0.0122	0.0164	-0.0122	0.0273
5	0.415	0.0013	0.415	0.0182	0.415	0.0281

6	0.4256	0.0014	0.4256	0.0187	0.4256	0.0283
7	-0.0818	0.0015	-0.0818	0.0195	-0.0818	0.0288
8	0.5466	0.0022	0.5466	0.0249	0.5466	0.0318
9	-0.1892	0.0023	-0.1892	0.0251	-0.1892	0.0319
10	0.6455	0.0032	0.6455	0.0306	0.6455	0.0355
11	-0.3422	0.0039	-0.3422	0.034	-0.3422	0.038
12	0.848	0.0059	0.848	0.0426	0.848	0.045
13	-0.7067	0.0099	-0.7067	0.056	-0.7067	0.0571
14	-0.839	0.0129	-0.839	0.064	-0.839	0.0647
15	-0.8829	0.014	-0.8829	0.0667	-0.8829	0.0673
16	-0.9552	0.0159	-0.9552	0.0712	-0.9552	0.0715
17	1.6721	0.0269	1.6721	0.0931	1.6721	0.093
18	2.4004	0.0589	2.4004	0.1376	2.4004	0.1399
19	4	0.173	4	0.2339	4	0.2646
20	-3.6755	0.1763	-3.6755	0.2361	-3.6755	0.2679

ขนาดตัวอย่างเท่ากับ 20 ค่าผิดปกติขนาดใหญ่ (6σ)

ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	l_1
1	0.221	0.0007	0.221	0.0127	0.221	0.0256
2	0.4256	0.0008	0.415	0.0138	0.415	0.0261
3	0.415	0.0008	0.4256	0.014	0.4256	0.0262
4	0.1017	0.0009	0.1017	0.0148	0.1017	0.0265
5	0.5466	0.0012	0.0181	0.0174	0.5466	0.0278

6	0.0181	0.0012	0.5466	0.0176	0.0181	0.0278
7	-0.0122	0.0013	-0.0122	0.0186	-0.0122	0.0283
8	0.6455	0.0017	-0.0818	0.0214	0.6455	0.0298
9	-0.0818	0.0017	0.6455	0.0215	-0.0818	0.0298
10	-0.1892	0.0024	-0.1892	0.0262	-0.1892	0.0326
11	0.848	0.0033	0.848	0.0309	0.848	0.0358
12	-0.3422	0.0038	-0.3422	0.0336	-0.3422	0.0377
13	-0.7067	0.0086	-0.7067	0.0519	-0.7067	0.0533
14	-0.839	0.0109	-0.839	0.0585	-0.839	0.0595
15	-0.8829	0.0117	-0.8829	0.0608	-0.8829	0.0616
16	-0.9552	0.0131	-0.9552	0.0644	-0.9552	0.0651
17	1.6721	0.0164	1.6721	0.0724	1.6721	0.0727
18	2.4004	0.0372	2.4004	0.1094	2.4004	0.1098
19	-3.6755	0.1281	-3.6755	0.202	-3.6755	0.2183
20	6	0.2671	6	0.2863	6	0.3648

ขนาดตัวอย่างเท่ากับ 50 ค่าผิดปกติขนาดเล็ก (3σ)

ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	l_1
1	0.314	0.0001	0.314	0.005	0.314	0.0099
2	0.3571	0.0001	0.3087	0.005	0.3087	0.0099
3	0.3087	0.0001	0.337	0.0053	0.337	0.0101
4	0.337	0.0001	0.3571	0.0057	0.3571	0.0102
5	0.1979	0.0002	0.1979	0.0062	0.1979	0.0105

6	0.3918	0.0002	0.3918	0.0066	0.3918	0.0107
7	0.1378	0.0003	0.1378	0.0082	0.1378	0.0116
8	0.4964	0.0004	0.4964	0.0106	0.4964	0.0131
9	0.5415	0.0005	0.5415	0.0125	0.5415	0.0145
10	0.5488	0.0006	0.5488	0.0129	0.5488	0.0148
11	0.5621	0.0006	0.5621	0.0135	0.5621	0.0152
12	0.6161	0.0008	-0.0234	0.0151	-0.0234	0.0166
13	-0.0234	0.0008	0.6161	0.0159	0.6161	0.0173
14	0.6674	0.0011	0.6674	0.0183	0.6674	0.0193
15	0.6842	0.0012	0.6842	0.019	0.6842	0.02
16	0.7414	0.0015	0.7414	0.0217	0.7414	0.0225
17	-0.2278	0.0019	-0.2278	0.0245	-0.2278	0.0252
18	0.8231	0.0021	0.8231	0.0254	0.8231	0.026
19	-0.2632	0.0022	-0.2632	0.0262	-0.2632	0.0267
20	-0.2833	0.0023	0.8545	0.0269	0.8545	0.0274
21	0.8545	0.0023	-0.2833	0.0271	-0.2833	0.0276
22	-0.2913	0.0024	-0.2913	0.0275	-0.2913	0.028
23	-0.3405	0.0028	-0.3405	0.0298	-0.3405	0.0302
24	-0.3464	0.0029	-0.3464	0.03	-0.3464	0.0305
25	-0.3676	0.003	-0.3676	0.031	-0.3676	0.0314
ลำดับที่	ข้อมูล	$\widetilde{K}$	ข้อมูล	$\widetilde{L}_1$	ข้อมูล	l_{L_1}
26	0.9828	0.0034	0.9828	0.0329	0.9828	0.0332
27	-0.4191	0.0035	-0.4132	0.0331	-0.4132	0.0335
28	-0.4132	0.0035	-0.4191	0.0334	-0.4191	0.0337
29	-0.4285	0.0036	-0.4285	0.0339	-0.4285	0.0342
30	1.0802	0.0044	1.0802	0.0374	1.0802	0.0376
31	-0.5657	0.0051	-0.5657	0.0403	-0.5657	0.0404
32	-0.5885	0.0054	-0.5885	0.0413	-0.5885	0.0415
33	1.1907	0.0057	1.1907	0.0426	1.1907	0.0427

34	1.2871	0.007	1.2871	0.0471	1.2871	0.0471
35	-0.7848	0.008	-0.7848	0.0505	-0.7848	0.0505
36	-0.8657	0.0092	-0.8657	0.0543	-0.8657	0.0543
37	-0.8845	0.0095	-0.8845	0.0551	-0.8845	0.0551
38	-0.9862	0.0112	-0.9862	0.0599	-0.9862	0.0599
39	1.5925	0.0118	1.5925	0.0613	1.5925	0.0613
40	1.6307	0.0125	1.6307	0.0631	1.6307	0.0631
41	1.7202	0.0142	1.7202	0.0673	1.7202	0.0674
42	-1.1647	0.0146	-1.1647	0.0682	-1.1647	0.0683
43	1.7481	0.0147	1.7481	0.0686	1.7481	0.0687
44	-1.5654	0.0237	-1.5654	0.0869	-1.5654	0.0875
45	-1.6471	0.0258	-1.6471	0.0907	-1.6471	0.0914
46	-1.6839	0.0268	-1.6839	0.0925	-1.6839	0.0932
47	-1.7092	0.0275	-1.7092	0.0936	-1.7092	0.0945
48	2.784	0.0428	2.784	0.1168	2.784	0.1193
49	3	0.0506	3	0.1269	3	0.1303
50	4.6447	0.1303	4.6447	0.2024	4.6447	0.2226

ขนาดตัวอย่างเท่ากับ 50 ค่าผิดปกติขนาดกลาง (4σ)

ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	L_1
1	0.314	0.0001	0.314	0.0049	0.314	0.0099
2	0.3571	0.0001	0.3087	0.0049	0.3087	0.0099
3	0.3087	0.0001	0.337	0.005	0.337	0.0099
4	0.3918	0.0001	0.3571	0.0053	0.3571	0.01
5	0.337	0.0001	0.3918	0.0059	0.3918	0.0104
6	0.1979	0.0002	0.1979	0.0066	0.1979	0.0107
7	0.4964	0.0003	0.1378	0.0087	0.1378	0.0119

8	0.1378	0.0003	0.4964	0.0094	0.4964	0.0123
9	0.5415	0.0004	0.5415	0.0112	0.5415	0.0136
10	0.5488	0.0005	0.5488	0.0115	0.5488	0.0138
11	0.5621	0.0005	0.5621	0.0121	0.5621	0.0142
12	0.6161	0.0007	0.6161	0.0144	0.6161	0.016
13	-0.0234	0.0008	-0.0234	0.0154	-0.0234	0.0168
14	0.6674	0.0009	0.6674	0.0167	0.6674	0.0179
15	0.6842	0.001	0.6842	0.0174	0.6842	0.0186
16	0.7414	0.0013	0.7414	0.0199	0.7414	0.0208
17	0.8231	0.0018	0.8231	0.0235	0.8231	0.0242
18	-0.2278	0.0019	-0.2278	0.0244	-0.2278	0.0251
19	0.8545	0.002	0.8545	0.0249	0.8545	0.0256
20	-0.2632	0.0022	-0.2632	0.026	-0.2632	0.0266
21	-0.2833	0.0023	-0.2833	0.0269	-0.2833	0.0274
22	-0.2913	0.0024	-0.2913	0.0273	-0.2913	0.0278
23	-0.3464	0.0028	-0.3405	0.0295	-0.3405	0.0299
24	-0.3405	0.0028	-0.3464	0.0297	-0.3464	0.0301
25	-0.3676	0.003	0.9828	0.0306	0.9828	0.031
ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	l_1
26	0.9828	0.003	-0.3676	0.0307	-0.3676	0.0311
27	-0.4191	0.0034	-0.4132	0.0327	-0.4132	0.033
28	-0.4132	0.0034	-0.4191	0.033	-0.4191	0.0333
29	-0.4285	0.0035	-0.4285	0.0334	-0.4285	0.0337
30	1.0802	0.0039	1.0802	0.035	1.0802	0.0353
31	-0.5657	0.0049	-0.5657	0.0395	-0.5657	0.0397
32	1.1907	0.005	1.1907	0.0399	1.1907	0.0401
33	-0.5885	0.0052	-0.5885	0.0405	-0.5885	0.0407
34	1.2871	0.0061	1.2871	0.0442	1.2871	0.0443
35	-0.7848	0.0076	-0.7848	0.0493	-0.7848	0.0493

36	-0.8657	0.0088	-0.8657	0.0529	-0.8657	0.0529
37	-0.8845	0.0091	-0.8845	0.0538	-0.8845	0.0538
38	1.5925	0.0105	1.5925	0.0579	1.5925	0.0579
39	-0.9862	0.0107	-0.9862	0.0583	-0.9862	0.0583
40	1.6307	0.0111	1.6307	0.0596	1.6307	0.0596
41	1.7202	0.0127	1.7202	0.0636	1.7202	0.0637
42	1.7481	0.0132	1.7481	0.0649	1.7481	0.0649
43	-1.1647	0.0138	-1.1647	0.0663	-1.1647	0.0664
44	-1.5654	0.0222	-1.5654	0.0843	-1.5654	0.0847
45	-1.6471	0.0242	-1.6471	0.0879	-1.6471	0.0885
46	-1.6839	0.0251	-1.6839	0.0896	-1.6839	0.0902
47	-1.7092	0.0258	-1.7092	0.0907	-1.7092	0.0914
48	2.784	0.0388	2.784	0.1112	2.784	0.1131
49	4	0.086	4	0.1651	4	0.1746
50	4.6447	0.1187	4.6447	0.1933	4.6447	0.2104

ขนาดตัวอย่างเท่ากับ 50 ค่าผิดปกติขนาดใหญ่ (6σ)

ลำดับที่	ข้อมูล	$\tilde{K}$	ข้อมูล	$\tilde{L}_1$	ข้อมูล	l_1
1	0.314	0.0001	0.3571	0.0049	0.314	0.0099
2	0.3571	0.0001	0.337	0.0049	0.3571	0.0099
3	0.3087	0.0001	0.314	0.005	0.337	0.0099
4	0.3918	0.0001	0.3087	0.0051	0.3087	0.01
5	0.337	0.0001	0.3918	0.0051	0.3918	0.01
6	0.4964	0.0002	0.4964	0.0074	0.4964	0.0111
7	0.1979	0.0002	0.1979	0.0074	0.1979	0.0112
8	0.5488	0.0003	0.5415	0.0088	0.5415	0.012
9	0.5621	0.0003	0.5488	0.0091	0.5488	0.0121

10	0.5415	0.0003	0.1378	0.0094	0.1378	0.0123
11	0.1378	0.0003	0.5621	0.0096	0.5621	0.0124
12	0.6161	0.0005	0.6161	0.0116	0.6161	0.0138
13	0.6674	0.0006	0.6674	0.0135	0.6674	0.0153
14	0.6842	0.0007	0.6842	0.0142	0.6842	0.0158
15	-0.0234	0.0008	-0.0234	0.0155	-0.0234	0.0169
16	0.7414	0.0009	0.7414	0.0164	0.7414	0.0177
17	0.8231	0.0013	0.8231	0.0197	0.8231	0.0206
18	0.8545	0.0014	0.8545	0.0209	0.8545	0.0217
19	-0.2278	0.0018	-0.2278	0.0236	-0.2278	0.0243
20	-0.2632	0.002	-0.2632	0.0251	-0.2632	0.0257
21	-0.2833	0.0021	-0.2833	0.0259	-0.2833	0.0264
22	0.9828	0.0022	0.9828	0.026	0.9828	0.0266
23	-0.2913	0.0022	-0.2913	0.0262	-0.2913	0.0267
24	-0.3405	0.0025	-0.3405	0.0282	-0.3405	0.0286
25	-0.3464	0.0026	-0.3464	0.0284	-0.3464	0.0289
ลำดับที่	ข้อมูล	$\widetilde{K}$	ข้อมูล	$\widetilde{L}_1$	ข้อมูล	l_1
26	-0.3676	0.0027	-0.3676	0.0292	-0.3676	0.0297
27	1.0802	0.0028	1.0802	0.0299	1.0802	0.0303
28	-0.4191	0.0031	-0.4132	0.0311	-0.4132	0.0315
29	-0.4132	0.0031	-0.4191	0.0313	-0.4191	0.0317
30	-0.4285	0.0032	-0.4285	0.0317	-0.4285	0.032
31	1.1907	0.0037	1.1907	0.0344	1.1907	0.0346
32	-0.5657	0.0044	-0.5657	0.0372	-0.5657	0.0374
33	-0.5885	0.0046	-0.5885	0.0381	-0.5885	0.0383
34	1.2871	0.0046	1.2871	0.0382	1.2871	0.0384
35	-0.7848	0.0066	-0.7848	0.046	-0.7848	0.0461
36	-0.8657	0.0076	-0.8657	0.0493	-0.8657	0.0493
37	-0.8845	0.0078	-0.8845	0.05	-0.8845	0.0501

38	1.5925	0.008	1.5925	0.0505	1.5925	0.0506
39	1.6307	0.0085	1.6307	0.0521	1.6307	0.0521
40	-0.9862	0.0092	-0.9862	0.0541	-0.9862	0.0541
41	1.7202	0.0097	1.7202	0.0557	1.7202	0.0557
42	1.7481	0.0101	1.7481	0.0568	1.7481	0.0568
43	-1.1647	0.0118	-1.1647	0.0613	-1.1647	0.0613
44	-1.5654	0.0188	-1.5654	0.0775	-1.5654	0.0777
45	-1.6471	0.0204	-1.6471	0.0807	-1.6471	0.0811
46	-1.6839	0.0212	-1.6839	0.0822	-1.6839	0.0826
47	-1.7092	0.0217	-1.7092	0.0832	-1.7092	0.0837
48	2.784	0.0304	2.784	0.0985	2.784	0.0996
49	4.6447	0.0943	4.6447	0.1727	4.6447	0.1839
50	6	0.1631	6	0.2258	6	0.2562

ภาคผนวก ค
โปรแกรมคอมพิวเตอร์ที่ใช้ในการวิจัย

โปรแกรมคอมพิวเตอร์ที่ใช้ในการวิจัย ผู้วิจัยใช้คำสั่งภาษา FORTRAN ซึ่งประกอบด้วยโปรแกรมหลักและโปรแกรมย่อย

โปรแกรมคอมพิวเตอร์นี้ เป็นการตรวจสอบค่าผิดปกติในข้อมูลตัวอย่างจากประชากรที่มีการแจกแจงแบบปกติที่ตัวอย่างขนาด 20 และกำหนดขนาดค่าผิดปกติเท่ากับ 6 ด้วยตัวสถิติ Generalized Extreme Studentized Deviate (GESD) , $\tilde{K}$, $\tilde{L}_1$ และ $\tilde{L}_1'$ โดยค่าผิดปกติที่ตรวจพบเป็นค่าผิดปกติที่เกิดจากการกำหนดและไม่ได้กำหนดของผู้วิจัย (แบบ A)

โปรแกรมหลักประกอบด้วยตัวแปร ดังนี้

DSEED	แทนจำนวนเริ่มต้น (Seed number) ในการสร้างเลขสุ่มของ IMSL LIBRARY
NR	แทนขนาดตัวอย่าง
NO1	แทนจำนวนค่าปกติที่คำนวณได้จากตัวสถิติ GESD
NO2	แทนจำนวนค่าปกติที่คำนวณได้จากตัวสถิติ $\tilde{K}$

NO3	แทนจำนวนค่าปกติที่คำนวณได้จากตัวสถิติ $\tilde{L}_1$
NO4	แทนจำนวนค่าปกติที่คำนวณได้จากตัวสถิติ L_1'
OUT1	แทนจำนวนค่าผิดปกติที่คำนวณได้จากตัวสถิติ GESD
OUT2	แทนจำนวนค่าผิดปกติที่คำนวณได้จากตัวสถิติ $\tilde{K}$
OUT3	แทนจำนวนค่าผิดปกติที่คำนวณได้จากตัวสถิติ $\tilde{L}_1$
OUT4	แทนจำนวนค่าผิดปกติที่คำนวณได้จากตัวสถิติ L_1'
R	แทนข้อมูลตัวอย่างที่สุ่มมาจากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน
XBAR	แทนค่าเฉลี่ยที่คำนวณจากข้อมูลทั้งหมด
XBARi	แทนค่าเฉลี่ยที่คำนวณจากข้อมูลที่ตัดค่าสังเกตออก 1 ค่า
MXBAR	แทนค่าเฉลี่ยที่คำนวณจากข้อมูลทั้งหมด ใช้สำหรับตัวสถิติ $\tilde{L}_1$ และ L_1'
MXBARi	แทนค่าเฉลี่ยที่คำนวณจากข้อมูลที่ตัดค่าสังเกตออก 1 ค่า ใช้สำหรับตัวสถิติ $\tilde{L}_1$ และ L_1'
SD	แทนส่วนเบี่ยงเบนมาตรฐานที่คำนวณจากข้อมูลทั้งหมด
SDi	แทนส่วนเบี่ยงเบนมาตรฐานที่คำนวณจากข้อมูลที่ตัดค่าสังเกตออก 1 ค่า

MSD	แทนส่วนเบี่ยงเบนมาตรฐานที่คำนวณจากข้อมูลทั้งหมด ใช้สำหรับตัวสถิติ $\tilde{L}_1$ และ L_1'
MSDi	แทนส่วนเบี่ยงเบนมาตรฐานที่คำนวณจากข้อมูลที่ตัดค่าสังเกตออก 1 ค่า ใช้สำหรับตัวสถิติ $\tilde{L}_1$ และ L_1'
ESD	แทนตัวสถิติ GESD
K	แทนตัวสถิติ $\tilde{K}$
L01	แทนตัวสถิติ $\tilde{L}_1$
L02	แทนตัวสถิติ L_1'
SK	แทนตัวสถิติ $\tilde{K}$ เมื่อถูกเรียงลำดับแล้ว
SL01	แทนตัวสถิติ $\tilde{L}_1$ เมื่อถูกเรียงลำดับแล้ว
SL02	แทนตัวสถิติ L_1' เมื่อถูกเรียงลำดับแล้ว
ALPHA	แทนระดับนัยสำคัญที่กำหนด สำหรับตัวสถิติ GESD เมื่อกำหนด $\alpha = 0.05$
LAMDA	แทนค่าวิกฤตในการทดสอบตัวสถิติ GESD
F	แทนฟังก์ชันที่ใช้คำนวณตัวสถิติ $\tilde{L}_1$

FF แทนฟังก์ชันที่ใช้คำนวณตัวสถิติ L_1

โปรแกรมหลักประกอบด้วยโปรแกรมย่อย ดังนี้

UMACH	:	โปรแกรมย่อยสำหรับกำหนดหน่วยข้อมูลเข้า (Input unit) และหน่วยผลลัพธ์ (Output unit) ในการเรียกใช้ IMSL LIBRARY
RNSET	:	โปรแกรมย่อยสำหรับกำหนดจำนวนเริ่มต้น (Seed number) ของ IMSL LIBRARY
RNNOA	:	โปรแกรมย่อยในการสร้างข้อมูลตัวอย่างที่มาจากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน ของ IMSL LIBRARY
Mean	:	โปรแกรมย่อยในการคำนวณค่าเฉลี่ย
Sdv	:	โปรแกรมย่อยในการคำนวณส่วนเบี่ยงเบนมาตรฐาน
MeanSdv_i	:	โปรแกรมย่อยในการคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน เมื่อตัดค่าสังเกตออก 1 ค่า
GESD	:	โปรแกรมย่อยในการคำนวณตัวสถิติ GESD
PERCEN_GESD:		โปรแกรมย่อยในการคำนวณค่าวิกฤตสำหรับตัวสถิติ GESD

TEST_GESD	:	โปรแกรมย่อยในการตรวจสอบค่าผิดปกติ สำหรับตัวสถิติ GESD
Kullback	:	โปรแกรมย่อยในการคำนวณตัวสถิติ $\tilde{K}$
L1_S	:	โปรแกรมย่อยในการคำนวณตัวสถิติ $\tilde{L}_1$
L1_i_S	:	โปรแกรมย่อยในการคำนวณตัวสถิติ $\acute{L}_1$
QDAGS	:	โปรแกรมย่อยสำหรับการอินทิเกรตของ IMSL LIBRARY เพื่อใช้คำนวณตัวสถิติ $\tilde{L}_1$ และ $\acute{L}_1$
SVRGN	:	โปรแกรมย่อยในการเรียงลำดับข้อมูลจากน้อยไปมาก ของ IMSL LIBRARY
Test_Bayes	:	โปรแกรมย่อยในการตรวจสอบค่าผิดปกติ สำหรับตัวสถิติที่อิงแนวคิดแบบเบย์

โปรแกรมหลัก

```
USE         MSIMSL
INTEGER     DSEED,NOUT,NR,H,C,T
INTEGER     NO1,NO2,NO3,NO4,OUT1,OUT2,OUT3,OUT4
PARAMETER   (C = 20,ALPHA = 0.05)
REAL        R(C),XBAR,XBARi(C),SD,SDi(C),K(C),L01(C),L02(C)
REAL        MXBAR,MXBARi,MSD,MSDi,SK(C),SL01(C),SL02(C)
REAL        AbsX(C),ESD,ALPHA,LAMDA
EXTERNAL    F,FF
COMMON      H,NF,MXBAR,MXBARi(C),MSD,MSDi(C)
```

```
CALL UMACH (2, NOUT)
DO    T = 1,2000
  NR    = C
  DSEED = 9999999 + (10*T**2)**2
  CALL RNSET (DSEED)
  CALL RNNOA (NR,R)
  R(NR) =  6
  CALL Mean(R,NR,XBAR)
  CALL Sdv(R,NR,XBAR,SD)
  CALL MeanSdv_i(R,NR,XBARi,SDi)
  CALL Kullback(NR,XBAR,XBARi,SD,K)
  CALL GESD(R,NR,XBAR,SD,AbsX,ESD)
  CALL PERCEN_GESD(ALPHA,NR,LAMDA)
  CALL TEST_GESD(ESD,LAMDA,NO1,OUT1)
  MXBAR  = XBAR
  MXBARi = XBARi
  MSD    = SD
  MSDi   = SDi
  NF     = NR

  CALL L1_S(NR,L01)
  CALL L1_i_S(NR,L02)
  CALL SVRGN (NR, K, SK)
  CALL SVRGN (NR, L01, SL01)
  CALL SVRGN (NR, L02, SL02)
  CALL Test_Bayes(NR,SK,NO2,OUT2)
  CALL Test_Bayes(NR,SL01,NO3,OUT3)
  CALL Test_Bayes(NR,SL02,NO4,OUT4)
END DO
```

```
      WRITE (NOUT,00011) NO1,OUT1
      WRITE (NOUT,00012) NO2,OUT2
      WRITE (NOUT,00013) NO3,OUT3
      WRITE (NOUT,00014) NO4,OUT4



00011 FORMAT ('NUMBER NO-OUTLIER AND OUTLIERS OF GESD  ::'2I6)
00012 FORMAT ('NUMBER NO-OUTLIER AND OUTLIERS OF KULLBACK ::', 2I6)
00013 FORMAT ('NUMBER NO-OUTLIER AND OUTLIERS OF L01 ::', 2I6)
00014 FORMAT ('NUMBER NO-OUTLIER AND OUTLIERS OF L02 ::', 2I6)

      END
```


จบโปรแกรมหลัก

โปรแกรมย่อยในการคำนวณค่าเฉลี่ย

```
      SUBROUTINE   Mean(X,N,XBAR)
      INTEGER      N
      REAL         SUMX ,X(N)

      SUMX = 0.0
      DO I = 1,N
          SUMX = SUMX+X(I)
      END DO
      XBAR = SUMX/N
```

```
RETURN
END
```

โปรแกรมย่อยในการคำนวณส่วนเบี่ยงเบนมาตรฐาน

```
SUBROUTINE    Sdv(X,N,XBAR,SD)
INTEGER       N
REAL          SUMXX ,X(N)

SUMXX = 0.0
DO I  = 1,N
   SUMXX = SUMXX+X(I)**2
END DO
SD = SQRT(SUMXX/(N-1) - (N*XBAR**2)/(N-1))
RETURN
END
```


โปรแกรมย่อยในการคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเมื่อตัดค่าสังเกตออก 1 ค่า

```
SUBROUTINE    MeanSdv_i(X,N,XBARi,SDi)
INTEGER           N
REAL          X(N),XBARi(N),SDi(N)

SUMi  = 0.0
SUMii = 0.0
DO I  = 1,N
```

```
      SUMi  = SUMi + X(I)
      SUMii = SUMii + X(I)**2
    END DO
    DO J  = 1,N
      SUMi  = SUMi - X(J)
      SUMii = SUMii - X(J)**2
      XBARi(J) = SUMi/(N-1)
      SDi(J)   = SQRT(SUMii/(N-2) - ((N-1)*XBARi(J)**2)/(N-2))
      SUMi  = SUMi + X(J)
      SUMii = SUMii + X(J)**2
    END DO
    RETURN
    END
```


โปรแกรมย่อยในการคำนวณตัวสถิติ $\tilde{K}$

```
    SUBROUTINE   Kullback(N,XBAR,XBARi,SD,K)
    INTEGER      N
    REAL         XBARi(N),K(N),NK

      NK = REAL(N)
      DO I = 1,N
        K(I) = (0.5*LOG(NK/(NK-1))) - 0.5/NK +
     +         (NK-1)*((XBAR - XBARi(I))**2)/(2*SD**2)
    END DO
    RETURN
    END
```

โปรแกรมย่อยในการคำนวณตัวสถิติ $\tilde{L}_1$

```
SUBROUTINE   L1_S(NF,L01)
INTEGER      NF,H
COMMON       H
REAL         A, ABS, B, ERRABS, ERREST, ERROR, ERRREL, EXACT,F
REAL         RESULT,L01(NF)
INTRINSIC    ABS
EXTERNAL     F

A = -10.0
B =  10.0
ERRABS = 0.0
ERRREL = 0.001
DO H = 1,NF

    CALL QDAGS (F, A, B, ERRABS, ERRREL, RESULT, ERREST)
    L01(H) = RESULT

EXACT = -4.0
ERROR = ABS(RESULT-EXACT)
END DO
RETURN
END
```


โปรแกรมย่อยในการคำนวณตัวสถิติ $\hat{L}_1$

```
SUBROUTINE   L1_i_S(NF,L02)
INTEGER      NF,H
COMMON       H
REAL         A, ABS, B, ERRABS, ERREST, ERROR, ERRREL, EXACT,FF
REAL         RESULT,L02(NF)
INTRINSIC    ABS
EXTERNAL     FF

A = -10.0
B =  10.0
ERRABS = 0.0
ERRREL = 0.001
DO H = 1,NF

   CALL QDAGS (FF, A, B, ERRABS, ERRREL, RESULT, ERREST)
   L02(H) = RESULT

EXACT = -4.0
ERROR = ABS(RESULT-EXACT)
END DO
RETURN
END
```


โปรแกรมย่อยในการคำนวณฟังก์ชันเพื่อนำไปคำนวณตัวสถิติ $\widetilde{L}_1$

```
REAL FUNCTION F(X)
INTEGER       H,NF
COMMON        H,NF,MXBAR,MXBARi(20),MSD
```

```
REAL          MXBAR,MXBARi,MSD,X
REAL          ABS,SQRT,EXP
INTRINSIC     ABS,SQRT,EXP

F = 0.5*ABS(SQRT((7.0*(NF-1))/(44*MSD**2))
+           *EXP(-(NF-1)*(X-MXBARi(H))**2/(2*MSD**2))  -
 +            SQRT((7.0*NF)/(44*MSD**2))
+           *EXP(-NF*(X-MXBAR)**2/(2*MSD**2)))

RETURN
END
```


โปรแกรมย่อยในการคำนวณฟังก์ชันเพื่อนำไปคำนวณตัวสถิติ L_1'

```
REAL FUNCTION FF(X)
INTEGER       H,NF
COMMON        H,NF,MXBAR,MXBARi(20),MSD,MSDi(20)
REAL          MXBAR,MXBARi,MSD,MSDi,X
REAL          ABS,SQRT,EXP
```

```
INTRINSIC     ABS,SQRT,EXP

FF = 0.5*ABS(SQRT((7.0*(NF-1))/(44*MSDi(H)**2))
+            *EXP(-(NF-1)*(X-MXBARi(H))**2/(2*MSDi(H)**2))   -
+             SQRT((7.0*NF)/(44*MSD**2))
+            *EXP(-NF*(X-MXBAR)**2/(2*MSD**2)))

RETURN
END
```


โปรแกรมย่อยในการตรวจสอบค่าผิดปกติ สำหรับตัวสถิติที่อิงแนวคิดแบบเบย์ ($\tilde{K}$ $\tilde{L}_1$ และ $\tilde{L}_1$)

```
SUBROUTINE   Test_Bayes(N,A,M1,M2)
INTEGER      N,MM,M1,M2
REAL         A(N)

IF ((A(N-1)-A(N-2).NE.0)) THEN
    MM = INT((A(N) - A(N-1))/(A(N-1) - A(N-2)))
    IF (MM.LT.2) THEN
       M1 = M1 + 1
```

```
    ELSEIF (MM.GE.2) THEN
       M2 = M2 + 1
    END IF
  ELSEIF ((A(N-1)-A(N-2).EQ.0)) THEN
        MM = INT(A(N) - A(N-1))
      IF (MM.LT.2) THEN
       M1 = M1 + 1
    ELSEIF (MM.GE.2) THEN
       M2 = M2 + 1
    END IF
END  IF
RETURN
END
```


โปรแกรมย่อยในการคำนวณตัวสถิติ GESD

```
SUBROUTINE   GESD(X,N,XBAR,SD,AbsX,ESD)
INTEGER      N
REAL         X(N),XBAR,SD,AbsX(N),MAXX(N)

DO I = 1,N
    AbsX(I) = ABS(X(I) - XBAR)
END DO

CALL SVRGN (N,AbsX,MAXX)
ESD = MAXX(N)/SD

RETURN
```

```
END
```

โปรแกรมย่อยในการคำนวณค่าวิกฤตสำหรับตัวสถิติ GESD


```
SUBROUTINE    PERCEN_GESD(ALPHA,N,LAMDA)
REAL          DF,P,T,TIN,LAMDA

P  = 1-(ALPHA/N)
DF = N-2
T  = TIN(P,DF)
LAMDA = (N-1)*T/SQRT((DF+T**2)*N)
RETURN
END
```

โปรแกรมย่อยในการตรวจสอบค่าผิดปกติ สำหรับตัวสถิติ GESD

```
SUBROUTINE    TEST_GESD(ESD,LAMDA,NO,OUT)
```

```
INTEGER      OUT,NO
REAL         ESD,LAMDA

IF (ESD.GE.LAMDA) THEN
   OUT = OUT + 1
ELSE
   NO = NO + 1
END IF
RETURN
END
```

ประวัติผู้วิจัย

ชื่อ - สกุล นายประสพชัย พสุนนท์

ประวัติการศึกษา

พ.ศ. 2541	ปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ มหาวิทยาลัยทักษิณ
พ.ศ. 2542	ศึกษาต่อระดับปริญญามหาบัณฑิต สาขาสถิติประยุกต์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ประวัติการทำงาน

พ.ศ. 2544 - ปัจจุบัน	อาจารย์สังกัดโปรแกรมวิชาคณิตศาสตร์และสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี สถาบันราชภัฏนครปฐม