K 42515003 : สาขาวิชาสถิติประยุกต์

คำสำคัญ : ค่าผิดปกติ / ความอ่อนไหวแบบเบย์ / ระยะทางคูลแบคก์ / ระยะทาง L_1

ประสพชัย พสุนนท์ : การตรวจสอบค่าผิดปกติตามแนวทางของเบย์ : ตัวสถิติระยะทางคูลแบคก์และ 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 $\hat{L}_1$ 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 $\tilde{K}$, $\tilde{L}_1$, $\hat{L}_1$ และ GESD ตรวจสอบค่าผิดปกติในข้อมูลตัวอย่างที่จำลองแบบขึ้น โดยทำซ้ำ 2,000 ครั้ง

ผลการวิจัยพบว่า

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

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

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 $\hat{L}_1$, 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 $\hat{L}_1$, $\tilde{L}_1$ and GESD respectively. With large magnitude of outliers , GESD can detect outlier most frequently , followed by $\tilde{K}$, $\hat{L}_1$ 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 $\hat{L}_1$, $\tilde{L}_1$ and GESD respectively. With medium and large magnitudes of outliers , GESD can detect outlier most frequently , followed by $\tilde{K}$, $\hat{L}_1$ 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 $\hat{L}_1$ 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.