ประสิทธิภาพของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนัก
ด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล

โดย
นางสาวกนกกาญจน์ รัตนไพบูลย์

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์
ภาควิชาคณิตศาสตร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2546

ISBN 974 – 464 – 428 – 1

ลิขสิทธิ์ของบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

THE EFFICIENCY OF THE MEAN ESTIMATOR USING WEIGHT BASED ON KERNEL DENSITY ESTIMATE

By

Kanokkarn Rattanaphiboon

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

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

MASTER OF SCIENCE

Department of Mathematics

Graduate School

SILPAKORN UNIVERSITY

2003

ISBN 974 – 464 – 428 – 1

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้วิทยานิพนธ์เรื่อง " ประสิทธิภาพของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล " เสนอโดย นางสาวกนกกาญจน์ รัตนไพบูลย์ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์

……..…….
(ผู้ช่วยศาสตรจารย์ ดร.จิราวรรณ คงคล้าย)
คณบดีบัณฑิตวิทยาลัย
วันที่ ….… เดือน ……..…… พ.ศ. ……..

ผู้ควบคุมวิทยานิพนธ์
รองศาสตราจารย์ ไพบูลย์ รัตนประเสริฐ

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

คณะกรรมการตรวจสอบวิทยานิพนธ์

……ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ปราณี นิลกรณ์)
…….. / ……..… / ……..

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

……กรรมการ
(รองศาสตราจารย์ วีรานันท์ พงศาภักดี)
…….. / ……..… / ……..

……กรรมการ
(อาจารย์ ดร.กมลชนก พานิชการ)
…….. / ……..… / ……..

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

K 44304201 : สาขาวิชาสถิติประยุกต์
คำสำคัญ : ค่าผิดปกติ / ตัวประมาณคงทน / การประมาณฟังก์ชันความหนาแน่น

กนกกาญจน์ รัตนไพบูลย์ : ประสิทธิภาพของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล (THE EFFICIENCY OF THE MEAN ESTIMATOR USING WEIGHT BASED ON KERNEL DENSITY ESTIMATE) อาจารย์ผู้ควบคุมวิทยานิพนธ์ : รศ. ไพบูลย์ รัตนประเสริฐ. 100 หน้า. ISBN 974 – 464 – 428 – 1

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการประมาณค่าเฉลี่ยของประชากร โดยใช้ค่าเฉลี่ยจากตัวอย่างที่ถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจากข้อมูลตัวอย่าง ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลถูกสร้างจากแนวความคิดที่จะลดอิทธิพลของค่าผิดปกติโดยการใช้ความหนาแน่นที่ประมาณขึ้นเป็นตัวถ่วงน้ำหนัก ในการศึกษาจะทำการเปรียบเทียบประสิทธิภาพของค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเทียบกับตัวประมาณค่าเฉลี่ยแบบอื่นๆ ได้แก่ ค่าเฉลี่ยจากตัวอย่าง Huber estimator Huber-type skipped mean estimator และ Three-part redescending estimator นอกจากนี้ได้ศึกษาคุณสมบัติของตัวประมาณค่าเฉลี่ยที่ใช้การถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลประกอบด้วย

จากการศึกษาโดยใช้การจำลองแบบ และกำหนดขนาดตัวอย่างเป็น 22 39 และ 100 พบว่า ในกรณีที่ข้อมูลไม่มีค่าผิดปกติตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยของประชากรเช่นเดียวกับค่าประมาณที่ได้จากค่าเฉลี่ยจากตัวอย่าง และดีกว่าค่าประมาณที่ได้จากตัวประมาณคงทนวิธีอื่นๆ นอกจากนี้เมื่อขนาดตัวอย่างเพิ่มขึ้นค่าประมาณจะเข้าใกล้ค่าเฉลี่ยของประชากรมากขึ้น สำหรับกรณีที่มีค่าผิดปกติในข้อมูล พบว่า ค่าเฉลี่ยของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะค่อยๆเบี่ยงเบนจากค่าเฉลี่ยของประชากรเช่นเดียวกับค่าเฉลี่ยจากตัวอย่างแต่เบี่ยงเบนช้ากว่า ในขณะที่ค่าเฉลี่ยของตัวประมาณคงทนไม่ค่อยมีการเปลี่ยนแปลงมากนัก

สำหรับการศึกษาคุณสมบัติคงทนของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล พบว่ารูปแบบของ Empirical Influence Function (EIF) ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเป็นแบบไม่มีขอบเขต เช่นเดียวกับ EIF ของค่าเฉลี่ยจากตัวอย่าง แต่จะมีลักษณะแบบค่อยๆ เพิ่มขึ้น โดยเพิ่มขึ้นช้ากว่า EIF ของค่าเฉลี่ยจากตัวอย่างและมี Breakdown point เป็น 0% ซึ่งถือว่าเป็นคุณสมบัติที่ไม่ดีสำหรับความคงทน อย่างไรก็ตามแม้ว่าตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะมีความคงทนไม่เท่ากับตัวประมาณคงทนที่ใช้กันอยู่ในปัจจุบัน แต่ถือว่าใช้ได้ดีในกรณีที่ไม่แน่ใจว่าข้อมูลที่กำลังศึกษามีข้อมูลผิดปกติหรือไม่

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ภาควิชาคณิตศาสตร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2546

ลายมือชื่อนักศึกษา .....

ลายมือชื่ออาจารย์ผู้ควบคุมวิทยานิพนธ์ .....

K 44304201 : MAJOR : APPLIED STATISTICS

KEY WORD : OUTLIERS / ROBUST ESTIMATORS / DENSITY ESTIMATION

KANOKKARN RATTANAPHIBOON : THE EFFICIENCY OF THE MEAN ESTIMATOR USING WEIGHT BASED ON KERNEL DENSITY ESTIMATE. THESIS ADVISOR : ASSO.PROF. PAIBOON RATANAPRASERT. 100 pp. ISBN 974 – 464 – 428 - 1.

The purpose of this research was to study the efficiency of mean estimator using weight based on kernel density estimate. The new estimator was created from the idea that attempts to reduce the effect of outliers by using weight based on density estimate. In the study, we compare the efficiency of the mean estimator using weight based on density estimate with those of the other mean estimators such as sample mean, Huber estimator, Huber type skipped-mean estimator and three-part redescending estimator. Another purpose was to study robustness properties of mean estimator using weight based on kernel density estimate.

The results of the simulation study, using sample size of 22 , 39 and 100, were as follow : in the case of normal data, the mean of the estimator using weight based on kernel density estimate is close to the population mean in the same way as sample mean and is closer to the population mean than those of the other robust estimators. When the sample size increased, the means of all estimators tend to be good estimators of the population mean. In the case of contaminated data, the mean of the mean estimator using weight based on kernel density estimate move away from the population mean in the same way as sample mean but relatively slower. While the mean of the robust estimators are not much affected.

The study of robustness properties of the mean estimator using weight based on kernel density estimate showed that : its empirical influence function was unbounded like that of the sample mean but it has less sharp and its breakdown point is zero the same as that of the sample mean which is not a good property for robustness. Although the mean estimator using weight based on density estimate is less robust than other robust estimators, it can be used when we are not sure whether the considering data contain some outliers or not.

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

Department of Mathematics Graduate School, Silpakorn University Academic Year 2003

Student's signature ……..

Thesis Advisor 's signature ……..

กิตติกรรมประกาศ

ในการศึกษาและเรียบเรียงวิทยานิพนธ์ครั้งนี้สำเร็จลุล่วงลงได้ด้วยการสนับสนุน การให้คำแนะนำอันมีค่า ตลอดจนกำลังใจที่ดียิ่งของบุคคลหลายท่าน ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ไพบูลย์ รัตนประเสริฐ กรรมการและอาจารย์ผู้ควบคุมวิทยานิพนธ์ ที่ได้ให้คำปรึกษา คำแนะนำ ข้อเสนอแนะต่างๆ และคอยตรวจแก้ข้อบกพร่องในการทำวิทยานิพนธ์มาโดยตลอด รวมถึงคณะกรรมการทุกท่านที่ได้กรุณาให้คำแนะนำเพิ่มเติมสำหรับการทำวิทยานิพนธ์ และขอกราบขอบพระคุณอาจารย์ทุกท่านในสาขาวิชาสถิติ ภาควิชาคณิตศาสตร์ มหาวิทยาลัยศิลปากรที่ได้ให้ความรู้และการฝึกฝนทักษะในการศึกษา

สุดท้ายนี้ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ ที่เคารพรัก ผู้ซึ่งคอยให้การสนับสนุน เป็นกำลังใจที่อบอุ่นและมีค่ามากสำหรับผู้วิจัย

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

สารบัญ

หน้า

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

สารบัญตาราง

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

สารบัญภาพ

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในการศึกษาค่าที่บอกลักษณะของประชากร ซึ่งเรียกว่า พารามิเตอร์(Parameter) เช่น ค่าเฉลี่ย มัธยฐาน และส่วนเบี่ยงเบนมาตรฐานของประชากรนั้น การที่จะหาค่าเหล่านี้ให้มีความถูกต้องแน่นอนจำเป็นจะต้องศึกษาจากทุกหน่วยในประชากร แต่ในทางปฏิบัติการศึกษาประชากรทั้งหมดเป็นไปได้ยากและไม่เป็นที่นิยมนัก เนื่องจากมีข้อจำกัดบางประการ เช่น ค่าใช้จ่าย ระยะเวลาที่ใช้ในการสำรวจข้อมูล หรืออาจไม่ทราบขอบเขตของประชากรที่แน่นอน เช่น ต้องการศึกษาผู้ป่วยโรคเอดส์ในประเทศไทย เราอาจทราบเพียงผู้ป่วยที่เข้ารับการรักษาตัวตามโรงพยาบาลต่างๆ แต่ไม่ครอบคลุมถึงผู้ป่วยที่ไม่ได้เข้ารับการรักษาที่โรงพยาบาลรวมทั้งผู้ติดเชื้อซึ่งเป็นกลุ่มที่เราไม่ทราบแน่นอน ดังนั้นขอบเขตของประชากรจึงกว้างยิ่งขึ้น การสำมะโนจะเสียเวลามากขึ้นและยังต้องเสี่ยงกับความคลาดเคลื่อนที่ไม่ใช่เนื่องมาจากการเลือกตัวอย่าง (Non-sampling error) ด้วย เพื่อหลีกเลี่ยงข้อจำกัดดังกล่าวจึงใช้การเลือกตัวแทนเพียงบางส่วนมาเป็นตัวอย่างแล้วใช้ข้อมูลจากตัวอย่างในการอนุมานค่าที่บอกลักษณะของประชากร ซึ่งมีเทคนิคการเลือกตัวแทนของประชากรซึ่งเรียกว่าเทคนิคการสุ่มตัวอย่าง (Sampling) อยู่หลายเทคนิค เช่น การสุ่มตัวอย่างแบบง่าย (Simple random sampling) เทคนิคการสุ่มตัวอย่างแบบมีระบบ (Systematic sampling) และเทคนิคการสุ่มตัวอย่างแบบมีชั้นภูมิ (Stratified sampling) เป็นต้น

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

โดยทั่วไปถ้าเราทราบการแจกแจงของประชากรเราสามารถที่จะหาตัวประมาณพารามิเตอร์ที่ใช้วัดตำแหน่งกลางของประชากรได้จากวิธีการต่างๆ เช่น วิธีภาวะน่าจะเป็นสูงสุด(Method of maximum likelihood) วิธีกำลังสองน้อยที่สุด(Least squares method) เป็นต้น พบว่าในหลายๆ การแจกแจงมี $\bar{X}$ เป็นตัวประมาณที่ดีเนื่องจากมีคุณสมบัติ UMVUE (Uniformly minimum variance unbiased estimator) คือ เป็นตัวประมาณค่าที่ไม่เอนเอียงและมีความแปรปรวนต่ำที่สุด ตัวอย่างเช่น ในกรณีของการแจกแจงที่มีพารามิเตอร์ตัวเดียว เช่น การแจกแจงแบบปัวซง (Poison distribution) พบว่ามี $\bar{X}$ เป็นตัวประมาณที่ดีของ λ (ซึ่งเป็นอัตราเฉลี่ยของการเกิดเหตุการณ์ “สำเร็จ” ต่อหน่วย) และในกรณีของการแจกแจงที่มีพารามิเตอร์ 2 ตัว เช่น การแจกแจงแบบปกติ (Normal distribution) ซึ่งมีพารามิเตอร์เป็น μ และ σ^2 ก็พบว่า $\bar{X}$ เป็นตัวประมาณที่ดีของ

μ ถึงแม้ว่า $\bar{X}$ จะเป็นตัวประมาณที่ดีและมีคุณสมบัติที่ดีหลายประการ แต่ในทางปฏิบัติข้อมูลที่เก็บมาอาจมีค่าผิดปกติ (Outlier) ซึ่งเป็นค่าของข้อมูลที่แตกต่างจากค่าของข้อมูลส่วนใหญ่ ค่าผิดปกติเพียงตัวเดียวก็อาจก่อให้เกิดความเสียหายอย่างมากคือทำให้ค่าประมาณของ μ เบี่ยงเบนไปจากค่าจริง เนื่องจาก $\bar{X}$ มีความไวต่อค่าผิดปกติดังนั้นจึงมีผู้เสนอกระบวนการคงทนขึ้น

ตัวสถิติคงทน (Robust statistics) ถูกเสนอขึ้นพร้อมกับความเป็นจริงที่ว่ามีข้อตกลง (Assumptions) มากมายที่ใช้สำหรับตัวสถิติ (เช่น Normality Linearity Independence) เพื่อให้การประมาณมีความน่าเชื่อถือ แต่ในทางปฏิบัติข้อมูลที่สำรวจได้อาจไม่เป็นไปตามข้อตกลงด้วยเหตุผลแรก คือ อาจพบว่ามีข้อมูลที่เป็นค่าผิดปกติ ซึ่งอาจก่อให้เกิดความเสียหายหากนำค่าของตัวสถิติที่มีความไวต่อค่าผิดปกติไปประมาณพารามิเตอร์ เนื่องจากอาจเกิดความผิดพลาดในการอนุมานประชากรและค่าประมาณที่ได้จะมีความน่าเชื่อถือลดลง จึงมีการคิดค้นวิธีการสำหรับใช้จัดการกับปัญหาค่าผิดปกตินี้ คือ การตั้งเกณฑ์เพื่อตรวจสอบค่าผิดปกติและเมื่อพบข้อมูลผิดปกติก็หาวิธีจัดการกับค่าผิดปกติดังกล่าวนั้นซึ่งมีได้หลายลักษณะ เช่น ตัดข้อมูลผิดปกติดังกล่าว หรือให้น้ำหนักน้อยๆ แก่ข้อมูลที่ผิดปกติ หรือเปลี่ยนวิธีการหาค่าประมาณค่ากลางโดยใช้ตัวสถิติตัวอื่นๆ ซึ่งวิธีการเหล่านี้ถูกพัฒนาขึ้นเป็นตัวสถิติคงทน นอกจากนี้การสมมติว่าการแจกแจงของค่าจากตัวอย่างสามารถประมาณได้ด้วยการแจกแจงแบบปกติเสมอ โดยอ้างจากทฤษฎีลิมิตสู่ส่วนกลาง (Central limit theorem : CLT) ก็อาจไม่เป็นจริงในทางปฏิบัติ ทั้งนี้ขึ้นอยู่กับขนาดตัวอย่างที่ใช้และประเภทของข้อมูลโดย Bessel (1818, quoted in Hampel 2001) Newcomb (1886, quoted in Hampel 2001) Jeffreys (1939, quoted in Hampel 2001) ได้แสดงว่าการแจกแจงของค่าจากตัวอย่างอาจคลาดเคลื่อนไปจากการแจกแจงแบบปกติ คือ อาจโด่งมากเกินไปหรือมีการกระจายสูง จึงอาจไม่เหมาะสมที่จะประมาณการแจกแจงของค่าจากตัวอย่างด้วยการแจกแจงแบบปกติ

การถกเถียงเกี่ยวกับความเหมาะสมของการปฏิเสธค่าผิดปกติเริ่มมีมานานแล้ว โดยพบครั้งแรกในงานวิจัยของ Daniel Bernoulli (1777, quoted in Hampel et al. 1986 : 34) และตามมาด้วย Bessel และ Baeyer (1876, quoted in Hampel et al. 1986 : 34) ขณะที่ Boscovich (1777, quoted in Hampel et al. 1986 : 34) พบว่าการปฏิเสธค่าผิดปกติจะเป็นไปตามแนวทางของแต่ละการวิเคราะห์ ซึ่งการปฏิเสธค่าผิดปกติวิธีหนึ่งที่ใช้กันมานานแล้ว คือ trimmed mean ซึ่ง Gergonne และ Stigler (1976, quoted in Hampel et al. 1986 : 34) รวมถึง Mendeleev (1895) พบว่านักวิจัยแต่ละคนก็มีความชำนาญในการใช้ trimmed mean ที่มีรูปแบบไม่เหมือนกัน นอกจากนี้มีการสร้างเกณฑ์ในการตัดสินมากมายที่จะแสดงว่าข้อมูลบางค่ามีความเบี่ยงเบนออกไปจากข้อมูลส่วนใหญ่ เกณฑ์ formal rejection คือ เกณฑ์ที่ใช้ในการตรวจสอบค่าผิดปกติซึ่งมี

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

หลายเกณฑ์ด้วยกัน ซึ่งถูกเสนอโดยนักสถิติหลายๆ คนปรากฏขึ้นครั้งแรกโดย Peirce (1852 : 161-163) และ Chauvenet (1863 : 469-566) ตามมาด้วย Stone (1868, quoted in Hampel et al. 1986 : 34) Wright (1884, quoted in Hampel et al. 1986:34) Irwin (1925, quoted in Hampel et al. 1986 : 34) Student (1927:151-164) Thompson (1935, quoted in Hampel et al. 1986:34) Pearson และ Chandra Sekar (1936, quoted in Hampel et al. 1986:34) และคนอื่นๆ อีกมากมาย

สำหรับวิธีการอื่นๆที่ถูกเสนอเพื่อใช้ในการจัดการกับปัญหาการเกิดค่าผิดปกติ เช่น Student(1927) ได้เสนอวิธีการสุ่มเพิ่มในกรณีของการเกิดค่าผิดปกติ โดยการสุ่มค่าสังเกตเพิ่ม เนื่องจากมีแนวคิดว่าตัวอย่างอาจมีขนาดเล็กเกินไปเมื่อเทียบกับประชากร ซึ่งการสุ่มตัวอย่างขนาดเล็กก็อาจพบว่ามีค่าซึ่งดูเสมือนเป็นค่าผิดปกติ สำหรับในกรณีอื่นๆที่ไม่ใช่การประมาณพารามิเตอร์ เช่น ในการวิเคราะห์การถดถอยแบบพหุ (Multiple regression analysis) ก็มีวิธีการที่ใช้จัดการกับปัญหาการเกิดค่าผิดปกติ เช่น วิธีการ leverage point และอื่นๆ นอกจากนี้เริ่มมีการพิจารณาตัวประมาณซึ่งให้น้ำหนักน้อยลงแก่ค่าสังเกตที่ผิดปกติ (Glaisher 1872 :73 ; E.J. Stone 1873 ; Edgeworth 1773 ; Newcomb 1886 ; Jeffreys 1932,1939, quoted in Hampel et al. 1986 :73) ความพยายามนี้เพื่อที่จะปรับค่าผิดปกติให้มีความเหมาะสมยิ่งขึ้น โดยการลดอิทธิพลของค่าผิดปกติให้สร้างความเสียหายน้อยลงดีกว่าที่จะแยกและกำจัดมันออกไปซึ่งเป็นจุดมุ่งหมายสำหรับทฤษฎีความคงทนสมัยใหม่

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ในปี 1940 และปี 1950 Tukey ได้แสดงให้เห็นถึงความไม่คงทนของค่าเฉลี่ยและได้เสนอทางเลือกอื่นคือการใช้ตัวสถิติคงทนซึ่งเป็นประโยชน์กว่า งานของเขาเป็นการสร้างการประมาณพารามิเตอร์ด้วยตัวสถิติคงทนและมีแนวทางที่แตกต่างจากผู้บุกเบิกก่อนหน้า (Tukey 1960 : 448-485) จากการพัฒนางานจำนวนมากพบวิธีที่จะสามารถจัดการกับปัญหาค่าผิดปกติ ซึ่งมีความน่าเชื่อถือมากกว่าและอยู่ภายใต้ทฤษฎีความคงทนเสนอโดย Huber(1964,1965,1968 , quoted in Huber 1981) และ Hampel(1968 , quoted in Hampel et al. 1986) โดยพวกเขาได้เสนอตัวสถิติคงทนซึ่งอยู่ในกลุ่มของ M-estimators ได้แก่ Huber estimator Huber – type skipped mean และ Three-part redescending estimator ซึ่งตัวประมาณเหล่านี้ถูกสร้างขึ้นจากแนวคิดที่จะให้น้ำหนักลดลงแก่ค่าสังเกตที่เป็นค่าผิดปกติเพื่อที่จะลดอิทธิพลของค่าผิดปกติให้มีความเหมาะสมมากยิ่งขึ้น โดยการกำหนดฟังก์ชันที่จะใช้เป็นตัวถ่วงน้ำหนักค่าสังเกตแต่ละค่า ซึ่งฟังก์ชันที่ใช้จะขึ้นอยู่กับแนวคิดของผู้สร้างแต่ละคนซึ่งอาจแตกต่างกันออกไป (ดูรายละเอียดเพิ่มเติมในบทที่ 2)

ในปัจจุบันนักสถิติหันมาให้ความสนใจกับกระบวนการทาง nonparametric มากขึ้น เนื่องจากไม่มีข้อตกลงหรือเงื่อนไขมากนัก โดยกระบวนการทาง nonparametric ส่วนใหญ่มักสนใจเพียงว่าข้อมูลมีการแจกแจงเหมือนกันหรือไม่โดยไม่สนใจรูปแบบของการแจกแจง ซึ่งบางครั้งอาจ

ไม่เพียงพอจึงมีการเสนอการประมาณความหนาแน่น (Density estimation) แบบ Nonparametric ขึ้น เพื่อให้ผู้วิเคราะห์สามารถเห็นภาพโดยรวมของรูปร่าง ลักษณะของการแจกแจง นอกจากนี้ยังเป็นประโยชน์ต่อกระบวนการวิเคราะห์ทางสถิติอีกด้วย วิธีการประมาณความหนาแน่นที่ง่ายที่สุด คือ ฮีสโตแกรม (Histogram) แต่วิธีนี้มีข้อเสีย คือ รูปร่างฮีสโตแกรมที่ได้จะขึ้นอยู่กับการเลือกจุดเริ่มต้นของรูปฮีสโตแกรมและช่วงกว้างของอันตรภาคชั้น ดังนั้นจึงมีผู้คิดค้นวิธีการประมาณความหนาแน่นวิธีอื่นๆ โดยวิธีการประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ในงานวิจัยนี้ถูกเสนอครั้งแรกโดย Fix และ Hodge (1951, quoted in Silverman 1986 : 1-2) ตามมาด้วย Silverman (1986) และ Nadaraya (1989, quoted in Allen 1997 : 145-150) นอกจากนี้ Rosenblatt (1956, quoted in Silverman 1986 : 6) ได้เขียนเอกสารเรื่องการประมาณความหนาแน่น Fix และ Hodge (1951, quoted in Silverman 1986 : 1-2) สร้างตัวประมาณความหนาแน่นแบบง่าย (Naive estimator) รวมทั้งเสนอการประมาณความหนาแน่นด้วยวิธีเนียร์เรสต์เนเบอร์(Nearest neighbor methods) Whittle (1958, quoted in Izenman 1991 : 205-204) สร้างตัวประมาณความหนาแน่นแบบฟังก์ชันถ่วงน้ำหนักในรูปทั่วไป (General weight function estimators) Cencov (1962, quoted in Izenman 1991 : 205-204) สร้างตัวประมาณความหนาแน่นโดยใช้อนุกรมออโธกอนอล (Orthogonal series estimators) Meisel (1973, quoted in Izenman 1991 : 205-204) ได้เสนอตัวประมาณความหนาแน่นเคอร์เนลโดยให้ window width แปรค่า ซึ่งเรียกว่าวิธี Variable kernel estimators Tapia และ Thompson (1978, quoted in Izenman 1991 : 205-204) สร้างตัวประมาณความหนาแน่นแบบภาวะน่าจะเป็นจำกัดสูงสุด (Maximum penalized likelihood estimators) Prakasa Rao's (1983, quoted in Silverman 1986 : 6) เสนอความคิดเกี่ยวกับการประมาณความหนาแน่นในเชิงทฤษฎี Devroye และ Gyorfi (1985, quoted in Silverman 1986 : 6) กลุ่มของ Wertz (1978, quoted in Silverman 1986 : 6) และ Delecroix (1983, quoted in Silverman 1986:6) อธิบายเทคนิคที่ใช้ในแต่ละวิธีการของการประมาณความหนาแน่น นอกจากนี้ Rosenblatt (1971, quoted in Silverman 1986 : 6) Fryer (1977, quoted in Silverman 1986 : 6) Wertz และ Schneider (1979, quoted in Silverman 1986 : 6) Bean และ Tsokos (1980, quoted in Silverman 1986 : 6) ได้ศึกษาเทคนิคอื่นๆที่เกี่ยวข้องกับการประมาณความหนาแน่นรวมทั้งการใช้งานกับข้อมูลจริง

เมื่อเราทราบความหนาแน่นที่จุดต่างๆ ของตัวแปรสุ่มก็เท่ากับทราบว่าตัวแปรสุ่มดังกล่าวมีค่าที่เป็นไปได้อะไรบ้างและค่าเหล่านั้นเกิดขึ้นด้วยความน่าจะเป็นเท่าไร ผู้วิจัยจึงสนใจที่จะนำความหนาแน่นมาปรับสำหรับการประมาณค่าพารามิเตอร์ โดยจะใช้แนวคิดเช่นเดียวกับตัวสถิติคงทนแต่จะใช้ค่าของความหนาแน่นที่ประมาณขึ้น ซึ่งเรียกว่าค่าประมาณความหนาแน่นเป็น

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตัวถ่วงน้ำหนัก นั่นคือใช้ค่าประมาณความหนาแน่นที่จุดต่างๆของข้อมูลตัวอย่างเป็นค่าถ่วงน้ำหนักแก่ข้อมูลแทนที่จะใช้การกำหนดตัวถ่วงน้ำหนักแบบคงที่ และเนื่องจากค่าผิดปกติเป็นค่าที่แตกต่างจากข้อมูลส่วนใหญ่ซึ่งจะอยู่บริเวณปลายหางของการแจกแจง ดังนั้นการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นจึงเสมือนเป็นการให้น้ำหนักน้อยๆ แก่ค่าผิดปกติที่อยู่ห่างออกไป วิธีการประมาณนี้จึงน่าจะเป็นทางเลือกอีกทางหนึ่งในการประมาณค่าพารามิเตอร์ที่น่าจะให้ความน่าเชื่อถือมากกว่าในกรณีที่ไม่แน่ใจว่าข้อมูลมีความผิดปกติหรือไม่ (ในกรณีที่แน่ใจว่าข้อมูลมีค่าผิดปกติการตัดข้อมูลที่ผิดปกติออกควรจะเป็นวิธีการที่ดีกว่า)

เพื่อเปรียบเทียบวิธีการประมาณที่เสนอใหม่ในการประมาณค่าเฉลี่ยของประชากร เราจะเปรียบเทียบร่วมกับวิธีอื่นๆ ที่นิยมใช้กันอีก 4 วิธี คือ ค่าเฉลี่ยจากตัวอย่าง ($\bar{X}$) Huber estimator Huber–type skipped mean estimator และ Three - part redescending estimator (หรืออาจเรียกว่า Hampel redescending estimator) โดยเราจะใช้ผลลัพธ์ร่วมกันจากการจำลองแบบจากการแจกแจงในหลายๆ ลักษณะเพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าต่างๆ ที่ได้กล่าวมาแล้ว ในการเปรียบเทียบจะวัดระยะห่างของค่าเฉลี่ยของค่าประมาณจากตัวประมาณดังกล่าวจากค่าเฉลี่ยของประชากร พร้อมทั้งหาความแปรปรวนและความคลาดเคลื่อนมาตรฐานของตัวประมาณแต่ละตัว

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบประสิทธิภาพของการประมาณค่าเฉลี่ยของประชากรที่ใช้การถ่วงน้ำหนักด้วยค่าประมาณที่ได้จากการประมาณความหนาแน่นแบบเคอร์เนลกับการประมาณค่าเฉลี่ยของประชากรแบบอื่นๆ ได้แก่ ค่าเฉลี่ยของตัวอย่าง ($\bar{X}$) Huber estimator Huber–type skipped mean estimator และ Three-part redescending estimator
2. เพื่อศึกษาสมบัติของตัวประมาณค่าเฉลี่ยที่ใช้การถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล

ขอบเขตของการวิจัย

1.ในการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของการประมาณค่าเฉลี่ยของประชากรที่ใช้การถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลกับการประมาณค่าเฉลี่ยของประชากรแบบอื่นๆ จะใช้การจำลองแบบ (Simulation) โดยกำหนดให้ตัวอย่างที่ใช้ในการทดลองเป็นตัวอย่างสุ่มจากการแจกแจงแบบปกติด้วยค่าเฉลี่ย μ และความแปรปรวน σ^2 ใช้สัญลักษณ์เป็น

$N(\mu, \sigma^2)$ ซึ่งในการศึกษาครั้งนี้จะกำหนดให้ $\mu = 0$ และ $\sigma^2 = 1$ นอกจากนั้นจะศึกษาตัวแปรสุ่มที่มาจากการแจกแจงที่มีลักษณะเบ้หรือการแจกแจงที่มีค่าผิดปกติเกิดขึ้น โดยผู้วิจัยจะใช้การแจกแจงที่ถือว่าเป็นการแจกแจงที่เกิดค่าผิดปกติ คือ การแจกแจงผสมระหว่างการแจกแจงแบบปกติ 2 การแจกแจงที่มีค่าเฉลี่ยและความแปรปรวนไม่เท่ากันผสมกัน ซึ่งเรียกการแจกแจงในลักษณะนี้ว่าการแจกแจงแบบ Contaminated normal ใช้สัญลักษณ์เป็น $CN(\mu_1, \sigma_1^2, p, \mu_2, \sigma_2^2)$ หมายความว่าการแจกแจงดังกล่าวเกิดจากข้อมูล 2 ส่วน คือ เกิดจากการแจกแจง $N(\mu_1, \sigma_1^2)$ ในสัดส่วน (1-p) และการแจกแจง $N(\mu_2, \sigma_2^2)$ ในสัดส่วน p เราจึงอาจกล่าวว่าการแจกแจง $N(\mu_1, \sigma_1^2)$ ปนกับ p % ของการแจกแจง $N(\mu_2, \sigma_2^2)$ โดยที่ p คือ สัดส่วนของตัวอย่างที่สุ่มจากการแจกแจงแบบ $N(\mu_2, \sigma_2^2)$ ที่นำไปผสมกับตัวอย่างที่สุ่มจากการแจกแจงแบบ $N(\mu_1, \sigma_1^2)$ ซึ่งในการศึกษาครั้งนี้จะกำหนดให้ ข้อมูลส่วนหนึ่ง มี $\mu_1 = 0$ และ $\sigma_1^2 = 1$ เสมอ ส่วนข้อมูลอีกส่วนหนึ่งมี $\mu_2 = 4$, 7 และ 10 โดยที่ $\sigma_2^2 = 1$ และมีสัดส่วนของการแจกแจงผิดปกติที่มาผสมกัน คือ p = 0.03, 0.05, 0.10, 0.20 และ 0.30

2. ขนาดตัวอย่างที่ใช้มี 3 ขนาด คือ n = 22 , 39 และ100 ตามลำดับ โดยอิงตามการศึกษาของ Allen(1997 : 145-150) ซึ่งถือว่าแทนตัวอย่างขนาดเล็ก กลาง และใหญ่ตามลำดับ

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

3. สำหรับ Huber estimator กำหนดเงื่อนไขค่าคงที่ b = 1.339 สำหรับ Huber–type skipped mean estimator ให้ r = 1.339 และสำหรับ Three-part redescending estimator ให้ a = 1.7, b = 3.4 , r = 8.5 ซึ่งเป็นค่าที่เหมาะสมจากการศึกษาของ Zhang (1996)

4. กำหนดฟังก์ชันเคอร์เนล (kernel function) ให้เป็นการแจกแจงแบบปกติมาตรฐาน (N(0,1)) เนื่องจากเป็นที่นิยมใช้ (ฟังก์ชันความหนาแน่นที่ประมาณโดยใช้ Kernel function เป็น N(0,1) จะเรียกว่า Gaussian kernel density estimate) และจากการศึกษาของ Devroye และ Gyorfi (1985) พบว่า Window width ที่ดีที่สุดสำหรับกรณีที่ประชากรมีการแจกแจงแบบปกติ และ kernel function ที่ใช้เป็น Gaussian kernel คือ Window width ที่มีค่าเท่ากับ $(2\sigma)/n^{1/5}$ แต่ในทางปฏิบัติเราไม่ทราบค่าของ σ^2 ดังนั้นในที่นี้จึงประมาณ σ^2 ด้วยความแปรปรวนจากตัวอย่าง ดังนั้น Window width ที่ใช้จึงอยู่ในรูป $(2s)/n^{1/5}$ เมื่อ s^2 คือความแปรปรวนของตัวอย่าง นอกจากนั้นยังศึกษา Window width อื่นอีก 2 ค่า คือกำหนดให้ Window width = $\frac{1}{2}s$ และ $\frac{1}{4}s$ ซึ่งเป็น $\frac{1}{2}$ เท่า และ $\frac{1}{4}$ เท่าของส่วนเบี่ยงเบนมาตรฐานจากตัวอย่าง

ประโยชน์ที่คาดว่าจะได้รับ

ได้ตัวประมาณค่าเฉลี่ยที่เป็นทางเลือกในกรณีที่ข้อมูลมีความผิดปกติ โดยให้การถ่วงน้ำหนักเป็นไปตามข้อมูลที่ได้มากกว่าจะเป็นการกำหนดรูปของการถ่วงน้ำหนักเฉพาะให้

นิยามและคำศัพท์เฉพาะ

1. ประชากร (Population) คือ สมาชิกทั้งหมดจากแหล่งของข้อมูลที่ต้องการเก็บรวบรวม สมาชิกดังกล่าวนี้จะต้องเป็นสมาชิกที่สามารถให้คำนิยามได้อย่างชัดเจนและสามารถชี้ระบุได้ โดยทั่วไปประชากรสามารถแบ่งออกเป็น 2 ประเภท คือ ประชากรที่มีจำนวนจำกัด (Finite population) และ ประชากรที่มีจำนวนไม่จำกัด (Infinite population) สำหรับประชากรที่มีจำนวนไม่จำกัดยังสามารถแบ่งได้เป็น 2 ประเภท คือ ประชากรที่มีจำนวนไม่จำกัดแบบนับได้ (Countably infinite population)และประชากรที่มีจำนวนไม่จำกัดแบบนับไม่ได้ (Uncountably infinite population)

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

กล่าวในแง่ของเซต ประชากร คือ เซตของสมาชิกทั้งหมดที่เป็นแหล่งของข้อมูลที่สนใจจะเก็บรวบรวม

2. ตัวอย่าง (Sample) คือ สมาชิกที่ถูกเลือกจากประชากรตามระเบียบวิธีที่กำหนดไว้

กล่าวในแง่ของเซต ตัวอย่าง คือ เซตย่อย (Subset) ที่มีสมาชิกซึ่งถูกเลือกมาจากประชากร

3. พารามิเตอร์ (Parameter) คือ ค่าคงที่ซึ่งอธิบายลักษณะของประชากร โดยคำนวณได้จากข้อมูลที่เก็บมาจากสมาชิกทั้งหมดของประชากร

4. ตัวสถิติ (Statistic) คือ ฟังก์ชันของตัวแปรสุ่มที่มีค่าเป็นค่าที่ใช้อธิบายลักษณะของตัวอย่างโดยคำนวณได้จากตัวอย่าง ใช้ประมาณพารามิเตอร์ของประชากร

5. ค่าผิดปกติ (Outlier) คือ ข้อมูลที่ไม่สอดคล้องกับรูปแบบของข้อมูลส่วนใหญ่ ไม่มีขอบเขตชัดเจน

6. ตัวสถิติคงทน (Robust statistic) คือ ตัวสถิติที่สร้างภายใต้ทฤษฎีความคงทนของกระบวนการทางสถิติ

7. ฟังก์ชันความหนาแน่น (Density function) คือ ฟังก์ชันความน่าจะเป็นของตัวแปรสุ่มแบบต่อเนื่องที่นิยามบน R = $(-\infty,\infty)$ แทนด้วยสัญลักษณ์ f(x) ซึ่งมีคุณสมบัติดังนี้ $f(x) \geq 0$ เมื่อ $-\infty < x < \infty$ และ $\int_{-\infty}^{\infty} f(x)dx = 1$

8. ค่าประมาณความหนาแน่นแบบเคอร์เนล(Kernel density estimating) คือ ความหนาแน่นที่ประมาณขึ้นโดยใช้วิธีการประมาณความหนาแน่นแบบเคอร์เนล ในการประมาณต้องกำหนด Kernel function และ Window width (ดูรายละเอียดเพิ่มเติมได้ในบทที่ 2 หน้า 22)

9. ฟังก์ชันการแจกแจง (Distribution function) คือ ฟังก์ชันความน่าจะเป็นสะสมของตัวแปรสุ่ม X แทนด้วยสัญลักษณ์ F(x) โดย $F(x) = P(X \leq x)$ เมื่อ $-\infty < x < \infty$

10. ประสิทธิภาพ (efficiency) คือตัวบ่งชี้ว่าตัวประมาณมีความแม่นยำมากน้อยเพียงใด โดยพิจารณาจากความใกล้เคียงของค่าคาดหวังของตัวประมาณค่าเทียบกับค่าพารามิเตอร์ที่ต้องการประมาณร่วมกับความแปรปรวนของตัวประมาณค่านั้นๆ

- ค่าคาดหวังของตัวประมาณเท่ากับพารามิเตอร์หรือไม่ $\left[E(\hat{\theta}) = \theta \right]$

11. Influence function คือ ฟังก์ชันที่ใช้อธิบายถึงอิทธิพลของค่าผิดปกติตัวเดียวต่อค่าของตัวประมาณ (สำหรับรายละเอียดเพิ่มเติมแสดงไว้ในบทที่ 2 หน้า 12)

12. Breakdown point คือ ขอบเขต หรือจำนวนของการเจือปนน้อยที่สุดของค่าผิดปกติที่เป็นสาเหตุให้ค่าของตัวประมาณเกิดการเปลี่ยนแปลง (สามารถดูรายละเอียดเพิ่มเติมได้ในบทที่ 2 หน้า 13)

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 ค่าเฉลี่ยจากตัวอย่าง

โดยทั่วไปในการศึกษาประชากรหนึ่งๆ สิ่งแรก คือ เราต้องการทราบพารามิเตอร์ของประชากร และการหาค่าเหล่านี้ให้มีความถูกต้องจำเป็นต้องศึกษาจากทุกหน่วยประชากร ซึ่งเป็นไปได้ยากในทางปฏิบัติ ดังนั้นจึงต้องเลือกตัวแทนเพียงบางส่วนมาเป็นตัวอย่างแล้วหาค่าที่ได้จากตัวอย่างเพื่อนำไปอนุมานประชากรต่อไป ค่าเฉลี่ยจากตัวอย่าง ($\bar{X}$) เป็นค่าหนึ่งที่นิยมใช้กันอย่างแพร่หลายในการประมาณค่าเฉลี่ยของประชากร เนื่องจาก $\bar{X}$ มีคุณสมบัติ UMVUE (Uniformly minimum variance unbiased estimator) คือ เป็นตัวประมาณค่าที่ไม่เอนเอียงและมีความแปรปรวนต่ำที่สุดในการประมาณค่าเฉลี่ยของประชากร โดยเราหาค่า $\bar{X}$ ได้ดังนี้

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

กำหนดให้ N แทนขนาดประชากร

n แทนขนาดตัวอย่าง

ให้ X_i เป็นตัวแปรสุ่มที่แทนค่าสังเกตจากหน่วยตัวอย่างที่ i และ x_i เป็นค่าของ X_i โดยที่ $i = 1, 2, \ldots, n$ ดังนั้น

ตัวประมาณค่าเฉลี่ยจากตัวอย่าง คือ $$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

ใช้ประมาณค่าเฉลี่ยของประชากร คือ $$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

และตัวสถิติความแปรปรวนของตัวอย่าง คือ $$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

ใช้ประมาณความแปรปรวนของประชากร คือ $$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

2.2 ตัวสถิติคงทน (Robust statistics)

กระบวนการคงทน เป็นกระบวนการทางสถิติที่อยู่ภายใต้แนวคิดหรือทฤษฎีความคงทน เพื่อที่จะหลีกเลี่ยงผลกระทบของการเบี่ยงเบนจากข้อสมมติของโมเดลอย่างมีระบบ และสิ่งหนึ่งซึ่งก่อให้เกิดการเบี่ยงเบนดังกล่าว คือ ค่าผิดปกติ ดังนั้นนักสถิติจึงได้คิดค้นวิธีการคงทนขึ้นเพื่อใช้ในการแก้ปัญหาการเกิดค่าผิดปกติ

เราอาจกล่าวได้ว่าข้อมูลที่มีคุณภาพสูงคือข้อมูลที่ปราศจากค่าผิดปกติ แต่ในความเป็นจริงเราอาจพบค่าผิดปกติในข้อมูลได้ ดังนั้นเมื่อสงสัยหรือแน่ใจว่ามีค่าผิดปกติเกิดขึ้น ควรที่จะหลีกเลี่ยงการใช้ตัวสถิติที่ไวต่อค่าผิดปกติ เช่น ค่าเฉลี่ยจากตัวอย่าง ($\bar{X}$) และหันมาเลือกใช้วิธีการที่อยู่ภายใต้ทฤษฎีความคงทนแทน ซึ่งหลักการของตัวสถิติคงทนอาจแบ่งได้เป็น 2 แบบ คือ แบบแรกจะใช้วิธีการตัดค่าผิดปกติออกโดยใช้กฎการปฏิเสธ (rejection rule) เช่น ปฏิเสธค่าสังเกตทุกตัวที่อยู่นอก [MED – 4MAD , MED + 4MAD] โดยที่ MED แทน มัธยฐาน ($med_i(x_i)$) และ MAD(median absolute deviation) $= median_i |x_i - MED|$ จากนั้นจึงคำนวณค่าประมาณของค่าเฉลี่ยจากค่าสังเกตที่เหลือ และแบบที่สอง คือ การลดอิทธิพลของค่าผิดปกติโดยการเลือกใช้ตัวประมาณคงทน เช่น การใช้ค่ามัธยฐาน (MED) ตัวประมาณ Hodges-Lehmann (H/L) ฯลฯ

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ตัวประมาณคงทนแต่ละตัวอาจให้ค่าประมาณที่ใกล้เคียงกัน หรือแตกต่างกันมากก็ได้ และตัวประมาณบางตัวอาจมีความคงทนมากกว่าตัวอื่นๆ (เช่น ค่ามัธยฐานมีความคงทนมากกว่า 10% trimmed -mean) แต่อย่างไรก็ตามตัวประมาณคงทนทุกตัวมีความสัมพันธ์สูงกับค่าเฉลี่ยจากตัวอย่างในกรณีที่ตัวอย่างมีความเป็นปกติ นั่นคือ ถ้าข้อมูลไม่มีค่าผิดปกติตัวประมาณทุกตัวจะให้ค่าประมาณใกล้เคียงกัน

เนื่องจากตัวประมาณคงทนแต่ละตัวมีความคงทนแตกต่างกัน ดังนั้นเพื่อความเข้าใจเราจำเป็นที่จะต้องเปรียบเทียบความคงทนของตัวประมาณค่าเหล่านี้โดยใช้เครื่องมือที่ใช้ในการวัดความคงทนซึ่งมีอยู่หลายวิธี เช่น การใช้ Influence function (IF) ซึ่งเป็นฟังก์ชันที่ใช้อธิบายถึงอิทธิพลของค่าผิดปกติเพียงตัวเดียวต่อค่าของตัวประมาณที่ใช้

Influence function (IF) ใช้ในการอธิบายผลกระทบของการเจือปนของค่าผิดปกติ x บนตัวประมาณ ซึ่งสามารถแสดงให้เห็นรูปร่างของพฤติกรรมการเปลี่ยนแปลงของค่าประมาณที่เกิดเนื่องจากค่าของ x เปลี่ยนไปเรื่อยๆ จาก $-\infty$ ไปยัง $+\infty$ (เดิมเรียกว่า Influence curve (Hampel 1968 , 1974) อย่างไรก็ตามทุกวันนี้เรามักใช้ชื่อ “Influence function” มากกว่า)

เราจะเริ่มการตรวจสอบความคงทนด้วย Influence function (IF) ซึ่งถูกเสนอโดย Hampel (1986 : 84) เพื่อที่จะตรวจสอบพฤติกรรมของฟังก์ชันค่าจริง

นิยาม Influence function (IF) ของตัวประมาณค่า T สำหรับฟังก์ชันการแจกแจง F คือ

$$IF(x;T,F) = \lim_{t\downarrow 0}\frac{T((1-t)F+t\Delta_x)-T(F)}{t}$$

สำหรับ $x \in \chi$ ที่ลิมิตหาค่าได้ (χ คือ เซตตัวอย่างที่เป็นไปได้ทั้งหมด(Sample space)) โดยที่ให้ Δ_x คือ ฟังก์ชันมวลความน่าจะเป็นที่จุด x เมื่อ T เป็นตัวประมาณค่า และ t เป็นสัดส่วนของค่าผิดปกติ

จากนิยามจะเห็นได้ว่า Influence function เป็นฟังก์ชันที่แสดงอัตราการเปลี่ยนแปลงของตัวประมาณค่าในกรณีที่มีค่าผิดปกติ และไม่มีค่าผิดปกติเทียบกับสัดส่วนของค่าผิดปกติ

เพื่อความเข้าใจเกี่ยวกับการหา Influence function เราจะใช้ตัวอย่างง่ายๆ เช่น ในกรณีของตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน ซึ่งมีค่าเฉลี่ยของตัวอย่างเป็นตัวประมาณของ Location parameter θ กำหนดให้ $F_\theta(x) = \Phi(x-\theta)$ เมื่อ Φ เป็นฟังก์ชันการแจกแจงสะสม (Cumulative distribution function : cdf) แบบปกติมาตรฐาน ที่มีฟังก์ชันความหนาแน่นเป็น $\phi(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2)$ ให้ $\theta_0 = 0$ ดังนั้น $F_{\theta_0}(x) = \Phi(x)$ โดยจะพิจารณาค่าเฉลี่ยเลขคณิต $T_n = (1/n)\sum_{i=1}^{n} X_i$ และสอดคล้องกับ $T(F) = E(T_n) = \int uf(u)du = \int udF(u)$ หาค่าได้จากนิยามของ IF จะได้ว่า

$$\begin{aligned} IF(x;T,\Phi) &= \lim_{t\downarrow 0}\frac{\int ud[(1-t)\Phi+t\Delta_x](u)-\int ud\Phi(u)}{t} \\ &= \lim_{t\downarrow 0}\frac{(1-t)\int ud\Phi(u)+t\int ud\Delta_x(u)-\int ud\Phi(u)}{t} \\ &= \lim_{t\downarrow 0}\frac{tx}{t} \end{aligned}$$

เพราะ $\int ud\Phi(u) = 0$ ดังนั้น

$$IF(x;T,\Phi) = x$$

นั่นคือเมื่อค่าผิดปกติมีการเปลี่ยนแปลงจาก $-\infty$ ไปยัง $+\infty$ ค่าเฉลี่ยของตัวอย่างซึ่งใช้เป็นตัวประมาณของประชากรจะมีการเปลี่ยนแปลงตามฟังก์ชัน $IF(x;T,\Phi) = x$ หรืออาจกล่าวได้ว่าเมื่อค่าผิดปกติ (x) เปลี่ยนไป 1 หน่วย ค่าประมาณของค่าเฉลี่ยก็จะเปลี่ยนไป 1 หน่วยด้วยเช่นกัน

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ตัวอย่างเช่น จากข้อมูล Cushny และ Peebles (1905) ซึ่งมีข้อมูลดังนี้ 0.0 0.8 1.0 1.2 1.3 1.3 1.4 1.8 2.4 4.6 เราสามารถใช้ Influence function ในการอธิบายถึงผลกระทบที่จะเกิดขึ้นกับค่าประมาณเมื่อค่าผิดปกติ(ในที่นี้คือ 4.6) มีค่าเปลี่ยนแปลงจาก $-\infty$ ไปยัง $+\infty$

นิยาม Empirical influence function(EIF)(Hampel et al. 1986 : 93) จากการเพิ่มค่าสังเกต คือ ถ้ามีตัวประมาณ $\{T_n;\ n \geq 1\}$ และตัวอย่าง $x_1,\ldots, x_{n-1}$ ขนาด n-1 แล้ว EIF ของตัวประมาณ คือ $T_n(x_1,\ldots, x_{n-1}, x)$ ซึ่งเป็นฟังก์ชันของ x ในทางกลับกันสามารถนิยาม EIF ได้โดยการแทนค่าสังเกต คือ เมื่อตัวอย่างประกอบด้วยค่าสังเกตจำนวน n ค่า เราสามารถแทนหนึ่งในค่าสังเกตจำนวน n ค่า(เรียกว่า x_n) ด้วยค่า x ใดๆ แล้วสร้าง EIF จาก $T_n(x_1,\ldots, x_{n-1}, x)$ หรือ $T_n(x_1,\ldots, x_n)$ สำหรับ $x = x_n$ เมื่อ x เป็นค่าผิดปกติเสมอ

ในที่นี้ Empirical influence function $EIF(x) = T_{10}(0.0 , 0.8 , \ldots , 2.4 , x)$ ซึ่งคือค่าของ Influence function จากตัวอย่าง เมื่อค่าผิดปกติซึ่งในที่นี้ คือ x มีค่าเปลี่ยนแปลงจาก $-\infty$ไปยัง $+\infty$ และ T คือ ตัวประมาณซึ่งเป็นฟังก์ชันของข้อมูลตัวอย่าง

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ภาพที่ 1 Empirical influence function ของค่าเฉลี่ย 10% trimmed mean และค่ามัธยฐาน

จากรูปจะเห็นได้ว่าค่าเฉลี่ยของตัวอย่างแตกต่างจากตัวประมาณอื่นๆ เนื่องจาก EIF ไม่มีขอบเขตแสดงว่าค่าเฉลี่ยของตัวอย่างไม่มีความคงทน จึงเปลี่ยนแปลงได้ง่ายเมื่อมีค่าผิดปกติเพียง

ตัวเดียว เครื่องมืออีกอย่างหนึ่งที่นิยมใช้วัดความคงทนคือ Breakdown point ของตัวประมาณซึ่งใช้ในการอธิบายว่าข้อมูลที่ศึกษาสามารถที่จะถูกเจือปนด้วยค่าผิดปกติได้มากเท่าไร ตัวประมาณของพารามิเตอร์โดยใช้ข้อมูลตัวอย่างจึงจะยังไม่มีผลกระทบ จากตัวอย่างของ Cushny และ Peebles 10 % trimmed mean และมัธยฐาน สามารถที่จะจัดการกับค่าผิดปกติเพียงตัวเดียวได้ อย่างไรก็ตาม ถ้าข้อมูลมีค่าผิดปกติ 2 ตัว 10% trimmed-mean จะใช้ไม่ได้ถ้าค่าผิดปกติอยู่บนข้างเดียวกัน เพราะในกรณี 10% trimmed-mean ค่าผิดปกติจะถูกตัดออกข้างละหนึ่งตัวเท่านั้น ดังนั้นในกรณีที่ค่าผิดปกติทั้งคู่ไม่ได้อยู่ข้างเดียวกันค่าประมาณจะยังคงใช้ได้ดี แต่ถ้ามีค่าผิดปกติ 3 ตัว (ในตัวอย่างจาก 10 ค่าสังเกต) ก็จะเป็นในทำนองเดียวกันกับ 20% trimmed-mean ด้วย นั่นคือค่าประมาณจะใช้ไม่ได้ถ้าค่าผิดปกติทั้ง 3 ตัวอยู่บนข้างเดียวกัน แต่จะยังใช้ได้หากมีค่าผิดปกติตัวหนึ่งที่ไม่ได้อยู่ข้างเดียวกัน ในขณะที่มัธยฐานจะยังคงใช้ได้เมื่อพบค่าผิดปกติถึง 4 ตัว

สำหรับเครื่องมืออีกอย่างหนึ่งที่ใช้ในการตรวจสอบความคงทนของตัวประมาณ คือ Breakdown point ซึ่งเป็นขอบเขต หรือจำนวนของการเจือปนที่น้อยที่สุดที่อาจเป็นสาเหตุให้ตัวประมาณเกิดการเปลี่ยนแปลง

นิยาม Breakdown point (BP) ของ Hodges(1967) มีดังนี้

สำหรับตัวอย่าง $X_1, X_2, \ldots, X_n$ Breakdown point ε_n^* ของตัวสถิติ T กำหนดโดย

$$\varepsilon_n^* = (T; x_1, x_2, ..., x_n) = \min\left\{\frac{m}{n};\ \max_{i_1,...,i_m} \sup_{y_1,...y_m} \left|T(z_1, z_2, ..., z_n)\right| < \infty\right\}$$

โดยที่ตัวอย่าง $z_1, z_2, \ldots, z_n$ ประกอบด้วยข้อมูลที่ได้จากการแทน $x_{i1}, x_{i2}, \ldots, x_{im}$ จำนวน m ค่าด้วยค่าใดๆ ของ $y_1, y_2, \ldots, y_m$ (Hampel et al. 1986:98)

จากนิยามจะเห็นได้ว่า Breakdown point ก็คือ สัดส่วนของจำนวนน้อยที่สุดของค่าผิดปกติที่จะทำให้ค่าประมาณเกิดการเปลี่ยนแปลง นั่นคือถ้ามีสัดส่วนของจำนวนของค่าผิดปกติน้อยกว่าหรือเท่ากับสัดส่วนดังกล่าวก็จะไม่ทำให้ค่าประมาณเปลี่ยนแปลงไป แต่ถ้ามีสัดส่วนของจำนวนค่าผิดปกติมากกว่าสัดส่วนดังกล่าวจะทำให้ค่าประมาณเกิดการเปลี่ยนแปลง นิยามเดิมของ Hodges(1967) จะเป็นกรณีของตัวอย่างขนาดจำกัด Hampel (1971) จึงเสนอรูปแบบเมื่อตัวอย่างมีขนาดใหญ่มากๆ (Asymptotic) ซึ่งจะเท่ากับ $\lim_{n\to\infty} \varepsilon^*_n = \varepsilon^*$

ปัจจุบันตัวสถิติคงทนได้ถูกคิดค้นขึ้นมากมายซึ่งสามารถแบ่งออกเป็นกลุ่มๆตามลักษณะของการสร้าง เช่น ตัวประมาณในกลุ่มของ M-estimators (มาจาก “Generalized maximum likelihood”) L-estimators (มาจาก “Linear combinations of order statistics”) R-estimators(เป็น

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

กลุ่มของตัวสถิติที่สร้างมาจาก Rank tests) ซึ่งปัจจุบันนักสถิติส่วนใหญ่ให้ความสนใจกับตัวประมาณที่ให้น้ำหนักน้อยลงแก่ค่าสังเกตที่ผิดปกติ จุดมุ่งหมายเพื่อที่จะปรับค่าผิดปกติให้เหมาะสมโดยให้ค่าผิดปกติสร้างความเสียหายน้อยลงดีกว่าที่จะแยกและกำจัดมันออกไป ในที่นี้จะใช้ตัวประมาณในกลุ่มของ M – estimators เพื่อเปรียบเทียบกับวิธีที่เสนอใหม่เนื่องจากในการทดสอบเปรียบเทียบต้องการที่จะเปรียบเทียบตัวประมาณที่อยู่ในรูปแบบค่าเฉลี่ย ดังนั้นจึงเสนอตัวประมาณที่อยู่ในรูปของการเฉลี่ยเช่นเดียวกัน คือ Huber estimator , Huber–type skipped mean estimator , Three-part redescending estimator (หรือ Hampel redescending estimator) ซึ่งเป็นตัวประมาณที่เป็นที่รู้จักและเป็นที่นิยมในกลุ่มของนักสถิติคงทน โดยตัวประมาณทั้ง 3 จะอยู่ในรูปของค่าเฉลี่ยถ่วงน้ำหนัก

2.2.1 M-estimators

เป็นที่ทราบดีว่าตัวประมาณภาวะน่าจะเป็นสูงสุด(MLE)ของพารามิเตอร์θ ถูกกำหนดโดย $T_n = T_n(X_1 , X_2 ,..., X_n)$ ซึ่งเป็นค่าที่ทำให้ Likelihood function ซึ่งก็คือ $\prod_{i=1}^{n} f(x_i;\theta)$ มีค่าสูงที่สุด (เมื่อ $f(x_i;\theta)$ คือ ฟังก์ชันความหนาแน่นของ X_i และ Likelihood function คือ ฟังก์ชันความหนาแน่นร่วมของ X_1 ถึง X_n เมื่อพิจารณาในเทอมของพารามิเตอร์ θ) หรือเทียบเท่ากับการทำให้สมการ

$$\sum_{i=1}^{n}[\ln f(x_i;\theta)] \quad = \quad \max_{T_n}!$$

เมื่อ ln คือ log ฐาน e และ $\max_{T_n}!$ คือ ค่าสูงที่สุดของ $\sum_{i=1}^{n}[\ln f(x_i;\theta)]$ จากค่า T_n ที่เป็นไปได้

หรือ $$\sum_{i=1}^{n}[-\ln f(x_i;\theta)] \quad = \quad \min_{T_n}! \qquad \text{.....2.2.1.1}$$

เมื่อ $\min_{T_n}!$ คือ ค่าน้อยที่สุดของ $\sum_{i=1}^{n}[-\ln f(x_i;\theta)]$ จากค่า T_n ที่เป็นไปได้

Huber (1964) ได้เสนอสมการใหม่เพื่อใช้ในการสร้าง M-estimators ดังนี้

$$\sum_{i=1}^{n}\rho(x_i,T_n) \quad = \quad \min_{T_n}! \qquad \text{.....2.2.1.2}$$

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

แก้สมการโดยการหาอนุพันธ์อันดับหนึ่งของ (2.2.1.2) เทียบกับพารามิเตอร์ เมื่อ ρ เป็นฟังก์ชันใดๆ ที่สามารถหาอนุพันธ์ได้ ให้ $\psi(x,\theta) = (\partial/\partial\theta)\rho(x,\theta)$ ดังนั้นค่าประมาณ T_n จะสอดคล้องกับสมการ

$$\sum_{i=1}^{n}\Psi(x_i,T_n) \quad = \quad 0 \qquad \text{.....2.2.1.3}$$

ตัวประมาณทุกตัวที่ได้มาจากการแก้สมการ (2.2.1.2) หรือ (2.2.1.3) จะถูกเรียกว่า M-estimator

แน่นอนว่าสมการ (2.2.1.2) และ (2.2.1.3) อาจไม่เหมือนกันเสมอไป แต่โดยทั่วไปสมการ (2.2.1.3) จะใช้ประโยชน์ได้มากกว่าเนื่องจากจะทำให้ง่ายขึ้นในการแก้สมการ ตัวประมาณที่ได้จะไม่เปลี่ยนแปลงเมื่อ Ψ ถูกคูณด้วยค่าคงที่ $r > 0$ [ค่าลบของ r จะสอดคล้องกับค่าสูงสุดใน(2.2.1.2)]

ข้อสังเกต MLE ทุกตัวคือ M-estimator (โดยที่ $\rho = -\ln f$ และ $\psi = -f'/f$) แต่ไม่ใช่ M-estimator ทุกตัวจะเป็น MLE

โดยทั่วไปตัวสถิติคงทนในกลุ่มของ M-estimators ถูกสร้างจากการกำหนดลักษณะของ influence function แต่เนื่องจาก IF เป็นสัดส่วน(แปรผันตรง)กับ ψ-function ดังนั้นจึงสามารถใช้การกำหนด ψ-function แทนได้ โดยผู้สร้างแต่ละคนจะกำหนดลักษณะของ ψ-function ที่ควรจะเป็นและสมเหตุสมผล เพื่อแทนลงในสมการ (2.2.1.3) และใช้ในการแก้สมการต่อไป ในที่นี้จะกล่าวถึง ψ-function ใน 3 รูปแบบซึ่งถูกเสนอขึ้นเพื่อใช้ในการสร้างตัวประมาณ Huber estimator Huber–type skipped mean estimator Three-part redescending estimators หรือ Hampel redescending estimator

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

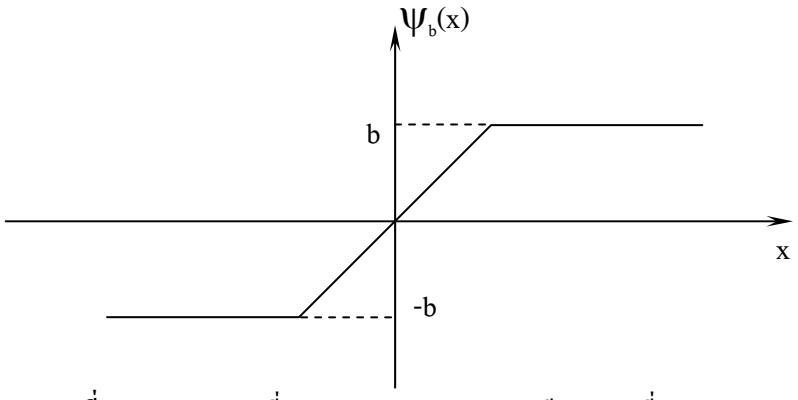
2.2.2 Huber estimator

The Huber estimator ถูกเสนอโดย Huber ในปี 1964 โดยกำหนด

$$\Psi_b(x) \quad = \quad \min\{b,\max\{x,-b\}\} \quad = \quad x\cdot\min\left(1,\frac{b}{|x|}\right)$$

หรืออาจเขียนให้อยู่ในรูปง่ายๆได้เป็น

$$\begin{aligned}\psi_b(x) \quad &= \quad x \qquad ; \ |x| \le b \\ &= \quad b \qquad ; \ |x| > b\end{aligned}$$



ภาพที่ 2 ψ -function ที่กำหนด Huber estimator มี cutoff ที่จุด b

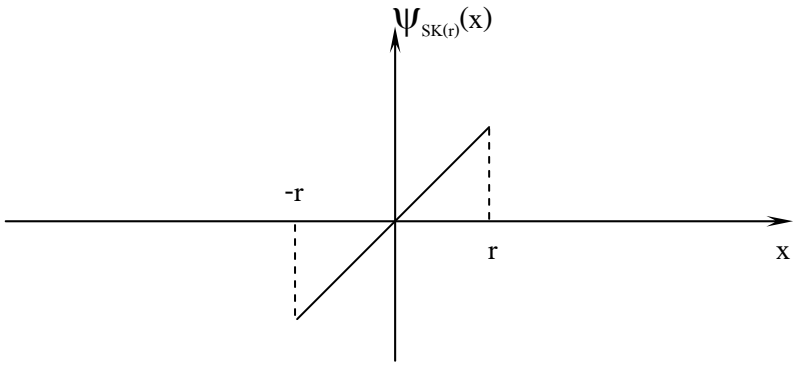
สำหรับ $0 < b < \infty$ จากภาพที่ 2 จะเห็นได้ว่า ψ -function มีขอบเขตแน่นอนในช่วง [-b , b] และจะมีค่าคงที่ถ้า |x| มากกว่า b โดยในกรณีที่ x มากกว่า b ψ -function จะมีค่าเท่ากับ b และกรณีที่ x น้อยกว่า b ψ -function จะมีค่าเท่ากับ –b

2.2.3 Huber-type skipped mean

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

Huber-type skipped mean estimator ถูกเสนอโดย Huber ในปี 1965 โดยกำหนด

$$\psi_{SK(r)}(x) = x \quad ; \ 0 \leq |x| \leq r$$
$$= 0 \quad ; \ r \leq |x|$$

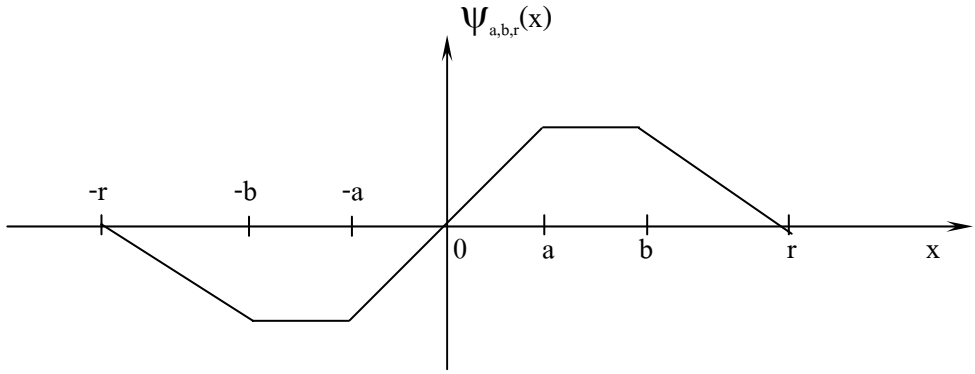


ภาพที่ 3 ψ -function ที่กำหนด Huber-type skipped mean

สำหรับ $0 < r < \infty$ จากภาพที่ 3 จะเห็นได้ว่า ψ -function มีขอบเขตแน่นอนในช่วง [-r , r] และจะมีค่าคงที่ถ้า |x| มากกว่า r คือ ψ -function จะมีค่าเป็นศูนย์

2.2.4 Three-part redescending estimator (Hampel estimator)

$$\begin{aligned}
\Psi_{a,b,r}(x) &= x && ; \ 0 \le |x| \le a \\
&= a\ sign(x) && ; \ a \le |x| \le b \\
&= a\frac{r-|x|}{r-b} sign(x) && ; \ b \le |x| \le r \\
&= 0 && ; \ r \le |x|
\end{aligned}$$



ภาพที่ 4 Ψ -function ที่กำหนด Three-part redescending M-estimator

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

เมื่อ $0 < a \le b < r < \infty$ ซึ่งบางครั้งถูกเรียกว่า "Hampels" เพราะตัวประมาณนี้ถูกเสนอโดย F. Hampel ใน Princeton Robustness Study (Andrews et al. 1972) จากภาพที่ 4 จะเห็นได้ว่า Ψ -function มีลักษณะค่อยๆ ลดลงเป็นช่วงๆ เมื่อ x มีค่าอยู่นอกช่วง [-a , a] โดยจะมีค่าคงที่ถ้า $a \le |x| \le b$ โดยจะมีค่าเป็น a [sign(x)] และจะมีค่าลดลงถ้า $b \le |x| \le c$ โดยจะมีค่าตามฟังก์ชัน $a\frac{r-|x|}{r-b} sign(x)$ และจะมีค่าเป็นศูนย์เมื่อ $r \le |x|$

ในการหาค่า T_n ของตัวประมาณในกลุ่ม M-estimators ทำได้โดยการแก้สมการซึ่งนิยามในรูป

$$\sum_{i=1}^{n} \Psi\left(\frac{x_i - T_n}{S_n}\right) = 0$$

การแก้สมการเพื่อหา T_n จะต้องผ่านกระบวนการทำซ้ำ ซึ่งอาจพบว่าไม่ได้มีเพียงคำตอบเดียว เนื่องจากการกำหนดค่าเริ่มต้นของ S_n ที่แตกต่างกัน สำหรับปัญหานี้อาจมีวิธีการแก้ไขหลายวิธี เช่น เลือกใช้ค่าที่ใกล้เคียงกับมัธยฐานมากที่สุดเป็นค่าประมาณค่ากลาง แต่ในที่นี้จะเสนอวิธีที่ง่ายที่สุด คือ การใช้ One-step M-estimators

ซึ่งกำหนดโดย

$$T_n = T_n^{(0)} + S_n^{(0)} \sum_{i=1}^{n} \Psi\left(\frac{x_i - T_n^{(0)}}{S_n^{(0)}} \right) \Bigg/ \sum_{i=1}^{n} \Psi'\left(\frac{x_i - T_n^{(0)}}{S_n^{(0)}} \right)$$

เมื่อ $S_n^{(0)}$ เป็น Robust estimator ของ Scale parameter ค่าของ $S_n^{(0)}$ ค่าแรกที่เหมาะสม กำหนดโดย

$$S_n^{(0)} \;=\; 1.483 MAD(x_i) \;=\; 1.483 med_i \mid x_i - MED \mid$$

โดย $MED = med_i(x_i)$ คือ ค่ามัธยฐานของข้อมูล x_i และ $med_i \mid x_i - MED \mid$ คือ มัธยฐานของข้อมูล $\mid x_i - MED \mid$ โดยที่ i มีค่าตั้งแต่ 1 ถึง n ดังนั้นเราสามารถหาค่า $S_n^{(0)}$ ซึ่งอยู่ในรูปของส่วนเบี่ยงเบนของค่าสังเกตจากมัธยฐานได้โดยการหาค่ามัธยฐานของชุดข้อมูล จากนั้นนำค่าสังเกตแต่ละตัวมาลบออกจากมัธยฐาน จากนั้นหาค่ากลางของค่าสัมบูรณ์ของค่าที่คำนวณได้ แล้วนำมาคูณกับค่าคงที่ 1.483

และ กำหนด $T_n^{(0)}$ ค่าแรกที่เหมาะสม ดังนี้

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

$$T_n^{(0)} \;=\; med\,(x_i)$$

Zhengyou Zhang (1996) ได้เขียนเกี่ยวกับตัวประมาณคงทนในกลุ่มของ M-estimators โดยเสนอตัวประมาณที่สร้างจากวิธีกำลังสองน้อยที่สุด ซึ่งปกติวิธีกำลังสองน้อยที่สุดพยายามที่จะหาค่าน้อยที่สุดของผลรวมของส่วนเหลือยกกำลังสองซึ่งจะไม่คงทนถ้าพบค่าผิดปกติในข้อมูล ตัวประมาณ M ถูกสร้างขึ้นเพื่อที่จะลดอิทธิพลของค่าผิดปกติโดยการแทนส่วนเหลือกำลังสอง (r^2 โดยที่ r แทนส่วนเหลือของค่าสังเกตที่ i จาก fitted value ของมัน) ด้วยฟังก์ชันใดๆ ที่สามารถหาอนุพันธ์ได้ และเรียกอนุพันธ์ของฟังก์ชันส่วนเหลือดังกล่าวว่า Influence function พร้อมทั้งเปรียบเทียบความคงทนของตัวประมาณที่ได้จาก Influence function ที่แตกต่างกัน

2.3 การประมาณความหนาแน่น (Density estimation)

ฟังก์ชันความหนาแน่นเป็นพื้นฐานที่สำคัญของการวิเคราะห์ทางสถิติ ทั้งนี้เพราะเมื่อทราบฟังก์ชันความหนาแน่นจะทำให้ทราบรูปร่างของการแจกแจงของประชากร และหาความน่าจะเป็นของการเกิดค่าต่างๆ ได้ มีความพยายามที่จะประมาณรูปกราฟของฟังก์ชันความหนาแน่น ซึ่งเรียกว่าวิธีการประมาณความหนาแน่น

โดยทั่วไปการประมาณความหนาแน่นอาจแบ่งได้เป็น 2 แบบด้วยกัน คือ การประมาณความหนาแน่นแบบ Parametric และการประมาณความหนาแน่นแบบ Nonparametric โดยการประมาณความหนาแน่นแบบ Parametric ทำได้ด้วยการสมมติว่าตัวอย่างถูกสุ่มมาจากประชากรในกลุ่มของการแจกแจงที่ทราบ ตัวอย่างเช่น สมมติว่าตัวอย่างถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 ดังนั้นฟังก์ชันความหนาแน่น f สามารถประมาณได้โดยการหาค่าประมาณของ μ และ σ^2 จากตัวอย่าง และแทนค่าประมาณดังกล่าวลงในสูตรฟังก์ชันความหนาแน่นแบบปกติ ในที่นี้เราจะไม่พิจารณาการประมาณความหนาแน่นแบบ Parametric แต่เราจะพิจารณาการประมาณความหนาแน่นแบบ Nonparametric เพียงอย่างเดียว เนื่องจากวิธีแบบ Nonparametric นี้มีข้อสมมติเกี่ยวกับการแจกแจงของประชากรที่ตัวอย่างถูกสุ่มมาที่มีความยืดหยุ่นมากกว่า และข้อมูลตัวอย่างควรจะถูกใช้ในการกำหนดค่าประมาณความหนาแน่นมากกว่าที่จะสมมติฟังก์ชันความหนาแน่น f ให้มีการแจกแจงอย่างใดอย่างหนึ่งในกลุ่มของการแจกแจงที่ทราบรูปแบบของการแจกแจง

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

นักสถิติหลายท่านมีความพยายามที่จะประมาณความหนาแน่นของประชากรจากชุดของตัวอย่างสุ่มที่สุ่มมาจากประชากร ซึ่งการประมาณความหนาแน่นที่จะกล่าวถึงนี้ถูกเสนอขึ้นครั้งแรกโดย Fix และ Hodges (1951 , quoted in Silverman 1986 : 2) เพื่อที่จะลดข้อสมมติเกี่ยวกับการแจกแจงของประชากร โดยทั่วไปการประมาณความหนาแน่นใช้เพื่อการตรวจสอบคุณสมบัติของเซตข้อมูลซึ่งสามารถบอกถึงลักษณะเด่นๆ ของข้อมูลที่เป็นไปได้ เช่น ความโด่งหรือการมีหลาย Mode ในข้อมูล และจะแสดงให้เห็นลักษณะของข้อมูลจริง

นอกจากนี้จุดมุ่งหมายที่สำคัญประการหนึ่งของวิธีการทางสถิติ คือ การนำเสนอข้อมูลกลับไปสู่ผู้ใช้ข้อมูลซึ่งอาจไม่ใช่นักคณิตศาสตร์ การประมาณความหนาแน่นถือเป็นแนวคิดหนึ่งสำหรับจุดมุ่งหมายนี้ และเพื่อความชัดเจนเราจะกำหนดสัญลักษณ์ที่ใช้เป็นมาตรฐานสำหรับหัวข้อนี้ โดยสมมติว่า $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มที่มาจากประชากรซึ่งฟังก์ชันความหนาแน่น f และกำหนดให้ $\hat{f}$ เป็นตัวประมาณความหนาแน่น เราจะเริ่มโดยการแนะนำวิธีการประมาณความหนาแน่นที่ง่ายและถูกใช้กันมานาน นั่นคือการประมาณโดยใช้ฮีสโตแกรม

2.3.1 การประมาณความหนาแน่นโดย ฮีสโตแกรม (Histogram)

วิธีการประมาณความหนาแน่นที่มีมานานและนิยมใช้กันอย่างแพร่หลาย คือ ฮีสโตแกรม (Histogram) การประมาณความหนาแน่นด้วยวิธีฮีสโตแกรมจำเป็นต้องกำหนดจุดกำเนิดเริ่มต้นและค่าช่วงของความกว้างเสียก่อน สมมติว่ามี x_0 เป็นจุดกำเนิดเริ่มต้น และ h เป็นขนาดความกว้างของช่วง(Bin width) สามารถสร้างช่วงความกว้างต่างๆ ของฮีสโตแกรมที่มีขนาด h ได้จาก $\left[(x_0 + mh) , (x_0 + (m+1)h) \right)$ สำหรับจำนวนเต็มบวกและลบ m ช่วงความกว้างถูกเลือกให้เป็นช่วงปิดทางซ้ายและช่วงเปิดทางขวา และค่าประมาณความหนาแน่น f ที่จุด x จากฮีสโตแกรมนิยามดังนี้

$$\hat{f}(x) = \frac{1}{nh} \text{ [จำนวนค่าสังเกตที่ตกอยู่ในช่วงเดียวกันกับ x]}$$

เมื่อ n เป็น ขนาดตัวอย่าง สังเกตว่าการสร้างฮีสโตแกรมจะต้องเลือกจุดเริ่มต้นและความกว้างของช่วง ซึ่งความกว้างของช่วงจะเป็นตัวควบคุมความราบเรียบของรูปฮีสโตแกรม

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

แนวคิดของการประมาณความหนาแน่นแบบนี้ คือ การตรึงจุดกำเนิดและความกว้างของช่วงต่างๆ ไว้ ความสูงของฮีสโตแกรมจะได้จากจำนวน x_i ที่ตกอยู่ในช่วงต่างๆ

สำหรับการแสดงและการตรวจสอบข้อมูล ฮีสโตแกรมถือเป็นเครื่องมือที่มีประโยชน์อย่างมากในการประมาณความหนาแน่นในกรณีของข้อมูลตัวแปรเดียว อย่างไรก็ตามเมื่อพิจารณาจากการประมาณความหนาแน่นด้วยวิธีฮีสโตแกรม พบว่าลักษณะรูปร่างของการแจกแจงโดยการใช้ฮีสโตแกรมมีความแตกต่างกันขึ้นอยู่กับการกำหนดจุดกำเนิดและความกว้างของช่วง นอกจากนี้ไม่สามารถหาอนุพันธ์ของฟังก์ชันประมาณความหนาแน่นที่จุดต่างๆ ได้ และสำหรับการแสดงรูปร่างของฮีสโตแกรมในข้อมูลตั้งแต่ 2 ตัวแปรขึ้นไปนั้นการสร้างยิ่งมีความลำบากขึ้น

จากจุดบกพร่องต่างๆ ของการประมาณความหนาแน่นด้วยวิธีฮีสโตแกรมจึงมีการพัฒนาวิธีการประมาณความหนาแน่นให้ดีขึ้น ในที่นี้ผู้วิจัยจึงขอยกตัวอย่างของการประมาณความหนาแน่นแบบง่าย (Naive estimator) เพื่อเป็นพื้นฐานสำหรับการประมาณความหนาแน่นแบบ Kernel ซึ่งเป็นวิธีที่ใช้ในการวิจัยครั้งนี้

2.3.2 ตัวประมาณความหนาแน่นแบบง่าย (The naive estimator)

วิธี Naive Estimator ใช้หลักการที่ว่าถ้าตัวแปรสุ่ม X มีฟังก์ชันความหนาแน่น f แล้ว

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x-h < X < x+h)$$

สำหรับค่า h ที่กำหนดใด ๆ เราสามารถประมาณ P(x-h < X < x+h) โดยใช้สัดส่วนของตัวอย่างที่อยู่ในช่วง (x-h < X < x+h) ดังนั้นตัวประมาณความหนาแน่น $\hat{f}$ ของ f ที่เป็นธรรมชาติ คือ การเลือกค่า h เล็กๆ และให้

$$\hat{f}(x) = \frac{1}{2hn} \text{ [จำนวนของค่าสังเกตที่ตกอยู่ในช่วง (x-h , x+h)]}$$

ซึ่งเรียกวิธีการประมาณความหนาแน่นด้วยวิธีการแบบนี้ว่า Naive estimator โดยที่ h เป็นความกว้างช่วง

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

เพื่อความเข้าใจวิธีของ Naive estimator ในการประมาณฟังก์ชันความหนาแน่น f มากขึ้น จะสมมติให้มีฟังก์ชันถ่วงน้ำหนัก w

โดยที่

$$w(x) = \begin{cases} 1/2 & ; \ |x| < 1 \\ 0 & ; \ \text{ที่อื่นๆ} \end{cases}$$

ดังนั้นวิธีการประมาณความหนาแน่นแบบง่าย สามารถเขียนได้ในรูปของฟังก์ชันถ่วงน้ำหนัก w ดังนี้

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h} \right)$$

ซึ่งเปรียบเสมือนกับการนำกล่องที่มีความกว้าง 2h และสูง $(2nh)^{-1}$ วางลงที่ค่าสังเกตแต่ละค่า และค่าประมาณความหนาแน่นที่จุด x ใดๆ จะได้จากส่วนของพื้นที่ซึ่งซ้อนทับกันขึ้นเป็นความสูงของค่าประมาณความหนาแน่นที่จุด x แต่การประมาณความหนาแน่นแบบง่ายนี้ยังคงมี

ปัญหาของการเลือกความกว้างช่วง h อยู่ ซึ่งจะมีผลต่อการประมาณความหนาแน่นเนื่องจากช่วงกว้างดังกล่าวนี้จะเป็นตัวควบคุมความราบเรียบของตัวประมาณความหนาแน่นดังนั้นการเลือกความกว้างช่วงที่ต่างกันจะทำให้รูปร่างของการแจกแจงที่ประมาณแตกต่างกันไปด้วย

จากแนวความคิดพื้นฐานของ 2 วิธีการที่กล่าวมาแล้ว ปัญหาที่สำคัญของการประมาณความหนาแน่น คือ ความไม่ต่อเนื่องและไม่สามารถหาอนุพันธ์ของฟังก์ชันที่ประมาณได้ ดังนั้นการประมาณความหนาแน่นแบบเคอร์เนลจึงได้รับการพัฒนาขึ้นเพื่อแก้ปัญหาดังกล่าว

2.3.3 การประมาณความหนาแน่นแบบเคอร์เนล (Kernel density estimation)

จากวิธีการประมาณความหนาแน่นแบบง่ายถ้าแทนฟังก์ชันถ่วงน้ำหนัก $w\left(\frac{x-X_i}{h}\right)$ ด้วย Kernel function K ซึ่งสอดคล้องกับเงื่อนไขดังนี้

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ซึ่งโดยทั่วไป K มักจะเป็นฟังก์ชันความหนาแน่นที่สมมาตร

จะได้วิธีการประมาณความหนาแน่นที่เรียกว่าวิธีการประมาณความหนาแน่นแบบเคอร์เนล (Kernel density function) ซึ่งนิยามดังนี้

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$$

เมื่อ K เป็น Kernel function และ h เป็น Window width

ตัวอย่าง จากข้อมูล Cushny และ Peebles(1905) ซึ่งมีข้อมูลดังนี้ 0.0 0.8 1.0 1.2 1.3 1.3 1.4 1.8 2.4 4.6 ต้องการประมาณความหนาแน่นที่ x = 4 เมื่อกำหนด Kernel function ให้เป็นการแจกแจงแบบปกติมาตรฐาน และกำหนดค่า Window width เป็น $2s/n^{1/5}$ โดยที่ s คือ ส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง ในที่นี้มีค่าเป็น 1.51

สามารถคำนวณค่า Window width ได้เป็น $(2\times1.51)/(10^{1/5}) = 1.9$

และคำนวณค่าของ $K\left(\frac{x-x_i}{h}\right)$ ที่ x = 4 ได้ดังตารางที่ 1

ตารางที่ 1 ค่าของ $K\left(\frac{x-x_i}{h}\right)$ ของค่าสังเกต x_i ต่างๆ ที่ x = 4 เมื่อ K(·) คือ Gaussian kernel

ค่าสังเกต (x_i)	$\left(\frac{x-x_i}{h}\right)$	$K\left(\frac{x-x_i}{h}\right)$
0.0	2.10526	0.043500
0.8	1.68421	0.096596
1.0	1.57895	0.114695
1.2	1.47368	0.134686
1.3	1.42105	0.145347
1.3	1.42105	0.145347
1.4	1.36842	0.156417
1.8	1.15789	0.204069
2.4	0.84211	0.279848
4.6	-0.31579	0.379538
รวม		1.70004

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

จะได้ $\sum_{i=1}^{10} K\left(\frac{4-x_i}{1.9}\right) = 1.70004$

จากนิยามข้างต้นค่าประมาณความหนาแน่นที่ x = 4 คำนวณได้ดังนี้

$$\hat{f}(4) = \frac{1}{10\times 1.9}[1.70004]$$

$$= 0.08947$$

ดังนั้นค่าประมาณความหนาแน่นที่ x = 4 เป็น 0.08947

ในขณะที่วิธีการประมาณความหนาแน่นแบบง่ายประมาณความหนาแน่นที่จุด x โดยเปรียบเสมือนการใช้กล่องสี่เหลี่ยมที่มีความกว้าง 2h และสูง $(2nh)^{-1}$ วางลงที่ค่าสังเกตแต่ละค่าและค่าประมาณความหนาแน่นที่จุด x ใดๆ จะได้จากส่วนของพื้นที่ซึ่งซ้อนทับกันขึ้นเป็นความสูงของค่าประมาณความหนาแน่นที่จุด x วิธีการประมาณความหนาแน่นแบบ Kernel จะเปรียบเสมือนกับการนำโค้ง (ตามรูปของ Kernel ที่ใช้) วางลงที่ค่าสังเกตแต่ละค่าและค่าประมาณความหนาแน่นที่จุด x ได้จากผลรวมของโค้งที่จุด x ดังกล่าว

ภาพที่ 5 การประมาณความหนาแน่นแบบ Kernel ที่แสดง Kernel ที่แต่ละจุด

ค่าของ h ในวิธีการประมาณความหนาแน่นแบบเคอร์เนล จะเรียกว่า Window width หรือ Bandwidth หรือบางครั้งเรียกว่า Smoothing parameter ทั้งนี้เพราะค่า h จะเป็นตัวกำหนดความราบเรียบของฟังก์ชันที่จะประมาณ วิธีประมาณความหนาแน่นแบบเคอร์เนลมีข้อดีหลายประการ คือ ฟังก์ชันที่ประมาณจะมีลักษณะต่อเนื่องและหาอนุพันธ์ได้ที่ทุกจุด

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

แต่อย่างไรก็ตามไม่อาจกล่าวว่าวิธีประมาณความหนาแน่นแบบเคอร์เนลดีที่สุดในบรรดาวิธีการประมาณความหนาแน่นแบบต่างๆ แต่เป็นวิธีที่นิยมใช้กันแพร่หลายมากที่สุดและได้มีผู้ศึกษาคุณสมบัติของวิธีประมาณความหนาแน่นไว้เป็นจำนวนมาก ดังนั้นผู้วิจัยจึงจะใช้วิธีการประมาณความหนาแน่นแบบเคอร์เนล เพื่อประมาณความหนาแน่นของประชากรสำหรับใช้ในงานวิจัยครั้งนี้เนื่องจากวิธีประมาณความหนาแน่นแบบเคอร์เนลเป็นวิธีการที่จะใช้ในการวิจัยครั้งนี้ จึงขอกล่าวถึงคุณสมบัติต่างๆของวิธีการนี้พอสังเขป

วิธีประมาณความหนาแน่นแบบ Kernel ในกรณีของตัวแปรเดียว

ให้ $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างที่เป็นอิสระและมีการแจกแจงเหมือนกัน จากการแจกแจงที่ต่อเนื่องที่มีตัวแปรเดียวและมีฟังก์ชันความหนาแน่น f ให้ $\hat{f}$ เป็นตัวประมาณความหนาแน่นของ f วิธีจะวัดว่าตัวประมาณมีประสิทธิภาพหรือไม่ ก็คือการหาความใกล้เคียงของตัวประมาณ $\hat{f}$ กับ f (Silverman 1986 : 34-44)

การวัดความแตกต่าง ระหว่าง $\hat{f}$ กับ f

ได้มีการศึกษาวิธีการวัดความแตกต่างของตัวประมาณความหนาแน่น $\hat{f}$ จากฟังก์ชันความหนาแน่นจริง f เมื่อพิจารณาการประมาณที่จุดๆเดียว วิธีการวัดที่เป็นธรรมชาติที่สุดก็คือ Mean square error ซึ่งใช้ชื่อย่อว่า MSE ซึ่งนิยามโดย

$$MSE_x(\hat{f}) = E\{\hat{f}(x) - f(x)\}^2$$

และโดยคุณสมบัติเบื้องต้นของ ค่าเฉลี่ยและความแปรปรวนจะได้

$$MSE_x(\hat{f}) = \{E\hat{f}(x) - f(x)\}^2 + Var\hat{f}(x) \quad \text{. . . . . 2.3.3.1}$$

ซึ่งเป็นผลบวกของกำลังสองของความเอนเอียงและความแปรปรวนที่จุด x เราจะเห็นว่าในหลายสาขาของทางสถิติจะมี Trade off ระหว่างความเอนเอียงและความแปรปรวนในรูป 2.3.3.1 กล่าวคือความเอนเอียงสามารถลดลงได้โดยการเพิ่มความแปรปรวนและในทำนองเดียวกันเมื่อลดความแปรปรวนก็จะเป็นการเพิ่มความเอนเอียง (สำหรับ MSE ที่คงที่)

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ในกรณีที่เป็นการประมาณฟังก์ชันที่ทุกๆจุด วิธีการที่นิยมกันมากที่สุดในการที่จะวัดความถูกต้องรวม(Global Accuracy) ของ $\hat{f}$ ในฐานะที่เป็นตัวประมาณค่าของ f ก็คือ Mean integrated square error (ใช้ชื่อย่อว่า MISE) ซึ่งนิยามโดย

$$MISE(\hat{f}) = E\int\{\hat{f}(x) - f(x)\}^2 dx$$

สังเกตว่าเนื่องจาก Integrand ไม่เป็นลบ ลำดับของ Integration และ Expectation สามารถที่จะสลับกันได้ซึ่งจะได้

$$\begin{aligned} MISE(\hat{f}) &= E\int\{\hat{f}(x) - f(x)\}^2 dx \\ &= \int MSE_x\ \hat{f}\, dx \\ &= \int\{E\hat{f}(x) - f(x)\}^2 dx + \int Var\hat{f}(x)\, dx \quad \text{..... 2.3.3.2} \end{aligned}$$

ซึ่งจะได้ว่า MISE เป็นผลบวกของ อินทิเกรตของกำลังสองของความเอนเอียง และอินทิเกรตของความแปรปรวน

สมมติว่า $\hat{f}$ เป็นตัวประมาณความหนาแน่นแบบถ่วงน้ำหนักในรูป

$$\hat{f}(x) = \frac{1}{n}\sum w(X_i, x)$$

เมื่อ w(•) เป็นฟังก์ชันถ่วงน้ำหนัก และ x เป็นจุดใดๆ บน R = (-∞, ∞)

จะได้ว่า $E\,\hat{f}(x) = \frac{1}{n}\sum Ew(X_i, x) = \int w(X_i, x) f(y)dy$

และเนื่องจาก X_i เป็นอิสระ ดังนั้น

$$\text{Var}\,\hat{f}(x) = \frac{1}{n}\left[Var \quad w(X_i, x) \right]$$

$$= \frac{1}{n}\left[\int w(y, x)^2 f(y)dy - \left\{ \int w(y, x) f(y)dy \right\}^2 \right]$$

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ซึ่ง MSE และ MISE สามารถหาได้โดยการแทน $E\,\hat{f}(x)$ และ $\text{Var}\,\hat{f}(x)$ ใน 2.3.3.2

คุณสมบัติที่น่าสนใจอย่างหนึ่งของ $E\,\hat{f}(x)$ ก็คือสำหรับ f ที่กำหนด ความเอนเอียง $E\,\hat{f}(x) - f(x)$ ไม่ขึ้นอยู่กับขนาดของตัวอย่างโดยตรงแต่ขึ้นอยู่กับฟังก์ชันถ่วงน้ำหนัก ซึ่งจุดนี้เป็นจุดที่สำคัญทั้งนี้ เพราะมันแสดงว่าการใช้ตัวอย่างที่เพิ่มมากขึ้นจะไม่เป็นตัวที่ลดความเอนเอียงตามลำพัง มันจำเป็นจะต้องปรับฟังก์ชันถ่วงน้ำหนักที่ใช้เพื่อที่จะได้การประมาณที่มีคุณสมบัติเป็น Asymptotically unbiased estimates

จากคุณสมบัติดังกล่าวนี้เมื่อนำมาใช้กับวิธีประมาณแบบ Kernel จะได้

$$E\,\hat{f}(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy \qquad \text{.....2.3.3.3}$$

และ $$nVar\hat{f}(x) = \int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left\{ \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy \right\}^2 \qquad \text{…….....2.3.3.4}$$

เมื่อ x เป็นจุดใดๆ บน R = (-∞, ∞) K(•) เป็นฟังก์ชัน Kernel และ h คือ Window width หรือ Smoothing parameter

ซึ่งเมื่อแทนค่า $E\,\hat{f}(x)$ และ $\text{Var}\,\hat{f}(x)$ ลงในสูตรของ MSE และ MISE ก็จะได้นิพจน์ที่แท้จริงของ MSE และ MISE สำหรับกรณีของวิธีประมาณฟังก์ชันความหนาแน่นแบบ Kernel

ซึ่งยกเว้นในกรณีเฉพาะมากๆ การคำนวณ $E\,\hat{f}(x)$ และ $Var\,\hat{f}(x)$ ไม่สามารถทำได้จึงจำเป็นที่จะต้องหาค่าประมาณของ $E\,\hat{f}(x)$ และ $Var\,\hat{f}(x)$ ต่อไป

จากรูปของ $E\,\hat{f}(x)$ นั่นคือ ค่าคาดคะเนของ $\hat{f}$ เป็น Smoothed version ของความหนาแน่นจริง ซึ่งหาได้จาก Convolving f ด้วย Kernel ที่ Scaled ด้วย Window width ซึ่งเป็นลักษณะของการประมาณความหนาแน่นเกือบจะทุกวิธี โดยที่การประมาณจะอยู่ในรูป

Smooth version of true density + random error .....……..…2.3.3.5

โดยที่ Smooth version of true density ขึ้นอยู่กับการกำหนดพารามิเตอร์ของวิธีการที่ใช้ แต่ไม่ได้ขึ้นโดยตรงกับขนาดตัวอย่าง

สำหรับกรณีเฉพาะกรณีหนึ่ง ที่สามารถจะหา $E\,\hat{f}(x)$ และ $Var\ \hat{f}(x)$ ในรูปที่ชัดเจนได้ก็คือในกรณีที่ Kernel คือการแจกแจงแบบปกติมาตรฐาน และความหนาแน่นจริงเป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 ซึ่งจะเห็นว่า $E\,\hat{f}$ ก็คือการแจกแจงแบบ $N(\mu, \sigma^2 + h^2)$ โดยการแทนค่า $E\,\hat{f}(x)$ และ $Var\,\hat{f}(x)$ ที่ใช้ Kernel คือการแจกแจงแบบปกติมาตรฐานลงในนิพจน์ของ $MSE_x(\hat{f})$ ก็จะได้นิพจน์ของ $MSE_x(\hat{f})$ ในรูปของผลรวมถ่วงน้ำหนักของฟังก์ชันความหนาแน่นของการแจกแจงแบบปกติ ซึ่งนิพจน์ดังกล่าวสามารถอินทิเกรตออกได้ดังนี้

$$(2\sqrt{\pi})MISE = \frac{1}{n}\left\{\frac{1}{h} - (\sigma^2 + h^2)^{-1/2}\right\} + \frac{1}{\sigma} + (\sigma^2 + h^2)^{-1/2} - 2\sqrt{2}(\sigma^2 + h^2)^{-1/2} \quad .....2.3.3.6$$

สามารถหา Optimum window width h ได้จากการทำให้สมการ 2.3.3.6 มีค่าน้อยที่สุด โดยการหาอนุพันธ์เทียบกับค่า h แล้วให้เท่ากับศูนย์

ความเอนเอียงและความแปรปรวน

ได้ชี้ให้เห็นแล้วว่าความเอนเอียงในการประมาณ $f(x)$ ไม่ได้ขึ้นโดยตรงกับขนาดของตัวอย่าง แต่ขึ้นอยู่กับ Window width h แน่นอนว่าถ้า h ถูกเลือกให้เป็นฟังก์ชันของ n ก็จะได้ว่าความเอนเอียงก็จะขึ้นอยู่กับ n โดยทางอ้อมเนื่องจากความเอนเอียงขึ้นอยู่กับ h

เราจะเขียน

$$bias_h(x) = E\hat{f}(x) - f(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy - f(x) \quad ..... 2.3.3.7$$

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ซึ่งความเอนเอียงก็จะขึ้นอยู่กับ Kernel K ด้วย แต่การขึ้นอยู่กับ Kernel K นี้ไม่ชัดเจน เราจะใช้ 2.3.3.7 เพื่อจะหานิพจน์ที่ประมาณความเอนเอียงนี้

โดยการเปลี่ยนตัวแปร y = x – ht และใช้ข้อสมมติที่ว่า K อินทิเกรตเป็น 1 จะได้

$$bias_h(x) = \int K(t) f(x - ht)\, dt - f(x)$$

$$= \int K(t) \{ f(x - ht)\, dt - f(x) \} dt$$

จากการกระจายอนุกรม Taylor ของ f(x-ht) จะได้

$$f(x - ht) = f(x) - htf'(x) + \frac{1}{2} h^2 t^2 f(x) + \ldots\ldots$$

ดังนั้นโดยข้อสมมติของ K ที่ว่า Kernel K เป็นฟังก์ชันสมมาตรที่สอดคล้องกับเงื่อนไข

$$\int K(t)dt = 1 \quad \int tK(t)dt = 0 \quad \text{และ} \quad \int t^2 K(t)dt = k_2 \neq 0$$ (เมื่อ k_2 เป็นค่าคงที่ใดๆ)

จะได้

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

$$bias_h(x) = -h f'(x) \int t K(t)\, dt + \frac{1}{2} h^2 f''(x) \int t^2 K(t) dt + \ldots\ldots$$

$$= \frac{1}{2} h^2 f''(x)\, k_2 + \text{higher-order terms in h}$$

ดังนั้นอินทิเกรตกำลังสองของความเอนเอียงในสูตร 2.3.3.2 สำหรับ Mean integrate square error จะคือ

$$\int bias_h(x)dx \approx \frac{1}{4} h^4 k_2^2 \int f''(x)\, dx \qquad \ldots\ldots\ldots\ldots\ldots\ldots 2.3.3.8$$

ต่อไปเรากลับไปที่ความแปรปรวน จาก 2.3.3.4 และ 2.3.3.3 เราได้

$$Var\hat{f}(x) = \frac{1}{n} \int \frac{1}{h^2} K\left(\frac{x-y}{h} \right)^2 f(y)dy - \frac{1}{n} \{ f(x) + bias_h(x) \}^2$$

$$\approx \frac{1}{nh} \int f(x - ht) K(t)^2 dt - \frac{1}{n} \{ f(x) + O(h^2) \}^2 \qquad \ldots\ldots\ldots\ldots 2.3.3.9$$

โดยการแทน y = x-ht ในอินทิกรัล และค่าประมาณ 2.3.3.9 สำหรับ bias สมมติว่า h เล็ก และ n มีขนาดใหญ่ โดยการกระจาย f(x-ht) ในรูปของอนุกรม Taylor จะได้

$$Var\hat{f}(x) \approx \frac{1}{nh}\int \{ f(x) - htf'(x) + ... \} K(t)^2 dt + O(n)^{-1}$$

$$= \frac{1}{nh} f(x) \int K(t)^2 dt + O(n)^{-1}$$

$$\approx \frac{1}{nh} f(x) \int K(t)^2 dt \qquad \text{..... 2.3.3.10}$$

เนื่องจาก f เป็นฟังก์ชันความหนาแน่น โดยการอินทิเกรตสมการ 2.3.3.10 ตามตัวแปร x จะได้การประมาณในรูปที่ง่ายคือ

$$\int Var\hat{f}(x) \approx \frac{1}{nh} \int K(t)^2 dt \qquad \text{.....……..... 2.3.3.11}$$

สมมติว่าเราต้องการเลือก h เพื่อที่จะทำให้ Mean integrate square error มีค่าน้อยที่สุดเท่าที่เป็นไปได้ เมื่อเปรียบเทียบค่าประมาณใน 2.3.3.8 กับ 2.3.3.11 สำหรับส่วนทั้งสองของ Mean integrate square error จะแสดงให้เห็นถึงปัญหาของการประมาณความหนาแน่นในความพยายามที่จะขจัดความเอนเอียงเราจำเป็นจะต้องใช้ค่า h ที่มีค่าเล็กซึ่งก็จะทำให้ Intergated varince มีค่ามาก ในขณะที่ถ้าเลือกค่า h ที่ใหญ่จะทำให้ความผันแปรสุ่มลดลง (ซึ่งสามารถวัดด้วยความแปรปรวน) จะเป็นการเพิ่ม Systematic error (หรือความเอนเอียง) ดังนั้นไม่ว่าวิธีการประมาณความหนาแน่นชนิดใดจะถูกใช้ การเลือก Smoothing parameter ก็จะเป็นการ trade off ระหว่าง random และ Systematic error

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

Window width ในอุดมคติ และ Kernel funtion

ค่าในอุดมคติของ h ในแง่ที่จะ การทำให้ค่าประมาณของ Mean integrated square error

$$\frac{1}{4} h^4 k_2^2 \int f''(x)^2 dx + \frac{1}{nh} \int K(t)^2 dt \qquad \text{..... 2.3.3.12}$$

มีค่าต่ำที่สุดสามารถหาได้ดังนี้

ถ้าให้ h_{opt} คือ Window width ที่ทำให้ Mean integrated square error มีค่าต่ำสุด โดยการหาอนุพันธ์ของ Mean integrated square error เทียบกับ h แล้วให้เท่ากับ 0 จะได้

$$h_{opt} = k_2^{-2/5} \left\{ \int K(t)^2 dt \right\}^{1/5} \left\{ \int f''(x)^2 dx \right\}^{-1/5} n^{-1/5} \quad \text{.....2.3.3.13}$$

และจากการศึกษาของ Devroye และ Gyorfi (1985) พบว่า Window width ที่ดีที่สุดสำหรับกรณีที่ประชากรมีการแจกแจงแบบปกติ และ Kernel function ที่ใช้เป็น Gaussian kernel คือ Window width ที่มีค่าเท่ากับ

$$h = \frac{2\sigma}{n^{1/5}}$$

เมื่อ σ เป็นส่วนเบี่ยงเบนมาตรฐานของตัวอย่างสุ่ม $X_1, X_2, \ldots, X_n$

สูตรในสมการ 2.3.3.13 สำหรับ Optimum window width ที่ได้ค่อนข้างจะน่าผิดหวัง ตรงที่ h_{opt} จะขึ้นอยู่กับฟังก์ชันความหนาแน่นซึ่งยังไม่ทราบค่าและกำลังจะประมาณอยู่ อย่างไรก็ตามเราก็ยังพอที่จะหาข้อสรุปที่เป็นประโยชน์ได้บ้าง ประการแรก Window width ในอุดมคติลู่เข้าสู่ 0 เมื่อขนาดตัวอย่างเพิ่มมากขึ้น แต่ในอัตราที่ช้ามาก ประการที่ 2 เนื่องจากเทอม $\int f''^2$ เป็นเสมือนตัววัดความรวดเร็วของการขึ้นลงในความหนาแน่น f จากสมการ 2.3.3.13 จะเห็นว่าค่าที่น้อยของ h จะเหมาะสมกับสำหรับฟังก์ชันความหนาแน่นที่มีการขึ้นๆ ลงๆ มาก ดังนั้นวิธีธรรมชาติที่ได้จาก 2.3.3.13 ก็คือเลือก h ที่อ้างอิงถึงฟังก์ชันความหนาแน่นมาตรฐาน ตัวอย่างเช่น การแจกแจงแบบปกติ

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

โดยการหาค่า h_{opt} จาก 2.3.3.13 แล้วแทนค่า h ลงในสมการ 2.3.3.12 จะได้ค่าประมาณของ Mean integrated square error เป็น $\frac{5}{4} C(K) \left\{ \int f''(x)^2 dx \right\}^{1/5} n^{-4/5}$ โดยที่ค่าคงที่ C(K) คือ

$$C(K) = k_2^{2/5} \left\{ \int K(t)^2 dt \right\}^{4/5} \quad \text{.....2.3.3.14}$$

จากการศึกษาสมการที่ 2.3.3.14 แสดงว่า ถ้าทุกสิ่งเท่ากัน เราควรเลือก Kernel K ที่มีค่า C(K) เล็กๆ ทั้งนี้เนื่องจากการเลือกค่า C(K) ดังกล่าวนี้จะทำให้ได้ค่า Mean integrated square error เล็กถ้าเราเลือก Smoothing parameter ถูกต้อง

สำหรับ Kernel ที่ตัวมันเองเป็นฟังก์ชันความหนาแน่น จะเป็น Kernel ที่ทำให้มั่นใจได้ว่าค่าประมาณ $\hat{f}$ จะมีค่าไม่เป็นลบ ถ้าค่าของ k_2 ไม่เท่ากับ 1 เราสามารถทำให้ k_2 เท่ากับ 1 ได้โดยการแทน Kernel ที่เปลี่ยนสเกลใหม่ในรูป $k_2^{1/2} K(k_2^{1/2} t)$ ซึ่งโดยการแทนค่าดังกล่าวจะไม่กระทบต่อค่าของ C(K)

ปัญหาของการทำให้ C(K) มีค่าน้อยที่สุดก็จะลดลงมาเป็นทำให้เป็นการหาค่าที่ต่ำสุดของ $\int K(t)^2 dt$ ที่สอดคล้องกับเงื่อนไขว่า $\int K(t)dt$ และ $\int t^2 K(t)dt$ ทั้งสองเทอมเท่ากับ 1 ซึ่ง Hodge และ Lehmann ได้พบว่า Kernel ที่สอดคล้องกับเงื่อนไขดังกล่าวคือ

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right) & ; \ -\sqrt{5} \le t \le \sqrt{5} \\ 0 & ; \text{ที่อื่นๆ} \end{cases} \qquad \text{.....2.3.3.15}$$

สัญลักษณ์ $K_e(t)$ ถูกใช้ในที่นี้ เพราะว่า Kernel ชนิดนี้ถูกเสนอในการประมาณความหนาแน่นโดย Epannechnikov(1969, quoted in Silverman : 42) ดังนั้น Kernel ชนิดนี้จึงถูกเรียกว่า Epannechnikov kernel

ตารางที่ 2 Kernel และ ประสิทธิภาพของ Kernel

Kernel	K(t)		Efficiency (exact and to 4 d.p.)
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right)$	สำหรับ $-\sqrt{5} \le t \le \sqrt{5}$	1
	0	ที่อื่นๆ	
Biweight	$\frac{15}{16}(1-t^2)^2$	สำหรับ $\lvert t \rvert < 1$	$\left(\frac{3087}{3125}\right)^{1/2} \approx 0.9939$
	0	ที่อื่นๆ	
Triangular	$1 - \lvert t \rvert$	สำหรับ $\lvert t \rvert < 1$	$\left(\frac{243}{250}\right)^{1/2} \approx 0.9859$
	0	ที่อื่นๆ	
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$		$\left(\frac{36\pi}{125}\right)^{1/2} \approx 0.9512$
Rectangular	$\frac{1}{2}$	สำหรับ $\lvert t \rvert < 1$	$\left(\frac{108}{125}\right)^{1/2} \approx 0.9295$
	0	ที่อื่นๆ	

ที่มา : Silverman, Density estimation for Statistics and data analysis (London : Chapman and Hall,1986), 43.

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ถ้านิยามประสิทธิภาพของ K เป็น

$$\text{eff(K)} = \left\{\frac{C(K_e)}{C(K)}\right\}^{5/4} \qquad \text{.....2.3.3.16}$$

$$= \frac{3}{5\sqrt{5}}\left\{\int t^2K(t)dt\right\}^{-1/2}\left\{\int K(t)^2dt\right\}^{-1} \qquad \text{.....2.3.3.17}$$

เหตุผลที่ใช้กำลังเป็น 5/4 ในสมการ 2.3.3.16 ก็เพราะ C(K) ติดอยู่ในรูปของฟังก์ชันกำลัง 4/5 ในตารางที่ 2 แสดง Kernel function และ ประสิทธิภาพของมัน เป็นที่น่าสังเกตว่าประสิทธิภาพที่ได้เมื่อเทียบกับ Epannechnikov kernel ซึ่งถือว่าเป็น Kernel ที่ดีที่สุดมีค่าใกล้เคียงกับ 1 แม้กระทั่ง Rectangular Kernel ที่ใช้ในกรณีการประมาณอย่างง่ายก็มีประสิทธิภาพใกล้กับ 0.93 ดังนั้นการเลือกชนิดของ Kernel ที่ใช้จึงไม่ใช่ปัญหาสำคัญทั้งนี้เพราะ Kernel เกือบจะมีประสิทธิภาพเท่ากัน ดังนั้นสิ่งที่ควรทำก็คือควรเลือก Kernel โดยการพิจารณาปัจจัยอื่นประกอบ เช่น อันดับของการหาอนุพันธ์ที่ต้องการ หรือ ความยากง่ายในการคำนวณ

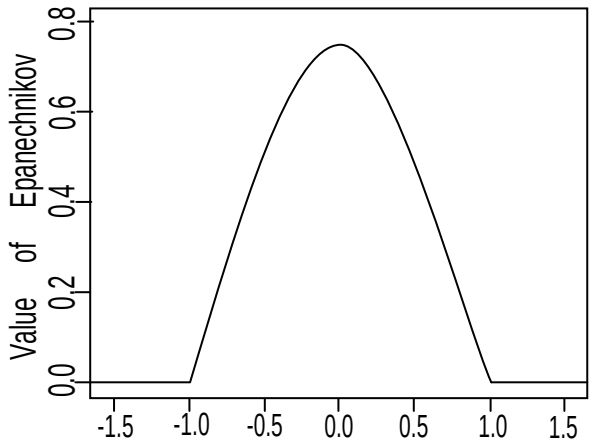
มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

การกำหนด Kernel Function

ในการประมาณความหนาแน่นจากวิธี Kernel density estimation ปัญหาต่อไป คือ การกำหนด Kernel function เป็นฟังก์ชันใช้สำหรับคลุม Window width ที่กำหนด โดยมีค่าของจุดตัวอย่างสุ่มที่สุ่มมาจากประชากรเป็นแกนกลาง ประสิทธิภาพของ Kernel function ที่ใช้สำหรับการวิธีประมาณค่าความหนาแน่นแบบ Kernel density estimator ได้มีผู้ศึกษากันมากพอสมควรแล้ว (ผู้สนใจอาจดูรายละเอียดได้จาก B.W. Silverman (1986)) มีข้อสรุปว่าการเลือก Kernel function ไม่ค่อยมีความสำคัญต่อประสิทธิภาพของวิธีการประมาณค่าความหนาแน่นแบบ Kernel มากนัก Kernel function ที่ดีที่สุดมีชื่อว่า Epanechnikov kernel ซึ่งมีรูปฟังก์ชัน ดังนี้

$$K(x) = \begin{cases} \frac{3}{4}\left(1-x^2\right), & \text{โดยที่} |x| < 1 \\ 0 \quad , & \text{ในที่อื่นๆ} \end{cases}$$

และมีรูปแบบการแจกแจง



เราอาจเรียก Kernel density estimate ที่ใช้ Kernel function อยู่ในรูป N(0,1) ว่า Gaussian kernel density estimate จากการศึกษาของ B.W. Silverman (1986) พบว่าการใช้ Normal density function เป็น Kernel function มีประสิทธิภาพประมาณ 0.9512 ของฟังก์ชัน Epanechnikov kernel แต่เนื่องจากเรารู้คุณสมบัติต่าง ๆ เช่น อนุพันธ์อันดับต่าง ๆ หรือ โมเมนต์ต่าง ๆ ดังนั้น Normal kernel จึงเป็นที่นิยมใช้กันมาก สำหรับการวิจัยในครั้งนี้จะใช้ Kernel function ในรูปของ Standard normal density function

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

การเลือก Window width หรือ Smoothing parameter

ปัญหาของการเลือกว่าจะให้ฟังก์ชันความหนาแน่นมีความราบเรียบมากน้อยเพียงใด เป็นปัญหาที่สำคัญในการประมาณความหนาแน่น ดังนั้นจะต้องไม่ลืมว่าการเลือก Smoothing parameter ที่เหมาะสมส่วนหนึ่งจะขึ้นอยู่กับวัตถุประสงค์ของการนำการประมาณความหนาแน่นไปใช้ ถ้าวัตถุประสงค์ของการประมาณความหนาแน่นเพื่อสำรวจวิเคราะห์ข้อมูลเพื่อที่จะหาตัวแบบและสมมติฐานที่เป็นไปได้ ซึ่งในกรณีเช่นนี้เราอาจเลือก Smoothing parameter แบบที่ขึ้นอยู่กับใจของผู้วิเคราะห์ แต่ถ้าใช้การประมาณความหนาแน่นเพื่อแสดงข้อสรุปเกี่ยวกับประชากรก็อาจจะให้ Undersmooth ได้บ้าง จากนั้นผู้ใช้ข้อมูลก็พอที่จะทำการปรับให้เรียบด้วยตาได้บ้าง

อย่างไรก็ตามในการประยุกต์ใช้ในหลายกรณีอาจต้องการเลือก Smoothing parameter ชนิดอัตโนมัติ ผู้ที่ไม่มีประสบการณ์อาจจะมีความรู้สึกสบายใจถ้าเลือกใช้ Smoothing parameter ที่เป็นชนิดอัตโนมัติ และวิธีเลือก Smoothing parameter แบบอัตโนมัตินี้สามารถใช้เป็นกรณีเริ่มต้นรายงานของนักวิทยาศาสตร์หรือกรณีที่ต้องการเปรียบเทียบผลของเขา อาจจะต้องการอ้างอิงถึงวิธี

ที่เป็นมาตรฐาน ถ้าการประมาณความหนาแน่นถูกใช้เป็นประจำบนตัวอย่างขนาดใหญ่หรือเป็นส่วนของวิธีการที่ใหญ่กว่าในกรณีดังกล่าววิธีที่เป็นอัตโนมัติก็จะมีความจำเป็น วิธีการเลือก Smoothing parameter มีอยู่ด้วยกันหลายวิธีซึ่งในที่นี้จะไม่ขอกล่าวถึงผู้สนใจอาจดูรายละเอียดได้จาก Silverman (1986)

ในการวิจัยครั้งนี้จะใช้การกำหนดค่าของ Window width จากการศึกษาของ Devroye และ Gyorfi (1985) ซึ่งพบว่า Window width ที่ดีที่สุดสำหรับกรณีที่ประชากรมีการแจกแจงแบบปกติ และ Kernel function ที่ใช้เป็น Gaussian kernel คือ Window width ที่มีค่าเท่ากับ

$$h = \frac{2\sigma}{n^{1/5}}$$

ดังนั้นเราจะประมาณ window width ด้วย

$$\hat{h} = \frac{2s}{n^{1/5}}$$

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

เมื่อ s เป็นส่วนเบี่ยงเบนมาตรฐานของตัวอย่างสุ่ม $X_1, X_2, \ldots, X_n$

นอกจากค่า Window width ดังกล่าวแล้วผู้วิจัยจะลองใช้ค่า Window width ค่าอื่น ๆ อีก เพื่อดูประสิทธิภาพในการประมาณว่าขึ้นอยู่กับค่า Window width อย่างไร ดังนั้นจึงกำหนดค่า Window width ค่าอื่นอีก 2 ค่า โดยจะศึกษาค่าของ Window width ที่เป็น 1/2 เท่า และ 1/4 เท่าของส่วนเบี่ยงเบนมาตรฐาน ซึ่งในการกำหนด Window width ให้อยู่ในรูปฟังก์ชันของส่วนเบี่ยงเบนมาตรฐานจากตัวอย่างด้วยเหตุผลที่ว่าค่าของ Window width ที่เหมาะสมในกรณีที่ตัวประมาณความหนาแน่นแบบเคอร์เนลที่ใช้มีฟังก์ชันความหนาแน่นแบบปกติ คือ ส่วนเบี่ยงเบนมาตรฐาน

Alan Julian Izenman (1991 : 205-224) ได้เสนองานวิจัยเรื่อง “การพัฒนาการประมาณความหนาแน่นแบบ Nonparametric” (*Recent Developments in Nonparametric Density Estimation*) โดยกล่าวว่าความก้าวหน้าในการคำนวณและเครื่องอำนวยความสะดวกในการคำนวณของนักสถิติมีผลกระทบต่องานวิจัยทางสถิติ นักสถิติได้พัฒนากระบวนการวิเคราะห์ข้อมูลแบบ Nonparametric ขึ้นมากมายทั้งในส่วนของทฤษฎีและการประยุกต์ เช่น งานวิจัยเกี่ยวกับการประมาณความหนาแน่นแบบ Nonparametric การวิเคราะห์การถดถอยแบบ Nonparametric การวิเคราะห์การจำแนก(Discrimination) แบบ Nonparametric และอื่นๆ ซึ่งเขาได้แสดงพัฒนาการ

ของการประมาณความหนาแน่นแบบ Nonparametric รวมถึงสิ่งที่อาจถูกมองข้ามไปเกี่ยวกับการประมาณความหนาแน่น โดยเริ่มจากวิธีการประมาณความหนาแน่นที่ง่ายที่สุด คือ ฮีสโตแกรม (Histogram) ตัวประมาณเคอร์เนล (Kernel estimators) และตัวประมาณความหนาแน่นโดยใช้อนุกรมออโธกอนอล (Orthogonal series estimators) ซึ่งเป็นที่รู้จักกันมาก รวมถึงการปรับความราบเรียบของความหนาแน่นที่ประมาณโดยวิธีต่างๆ และตัวประมาณความหนาแน่นแบบอื่นๆ

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

บทที่ 3

วิธีดำเนินงานวิจัย

การศึกษาในงานวิจัยนี้อาจแบ่งได้เป็น 2 การศึกษา ดังนี้

1. การศึกษาแรกเป็นการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าเฉลี่ยที่ได้จากวิธีต่างๆ ได้แก่ ตัวประมาณค่าเฉลี่ยที่ใช้การถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล ตัวประมาณค่าเฉลี่ยที่ใช้ค่าเฉลี่ยจากตัวอย่าง และตัวประมาณค่าเฉลี่ยจากวิธีสถิติคงทนซึ่งจะใช้ ตัวประมาณ 3 ตัวในกลุ่มของ M estimator นั่นคือ Huber estimator Huber–type skipped mean estimator และ Three-part redescending estimator โดยจะใช้โปรแกรม Fortran power station 4.0 เพื่อช่วยในการสุ่มตัวอย่างและคำนวณค่าต่างๆ ที่ใช้ในงานวิจัย

สำหรับการศึกษานี้อาจแบ่งได้เป็น 2 กรณีด้วยกัน

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

1) กรณีแรกเป็นการเปรียบเทียบประสิทธิภาพของตัวประมาณต่างๆ ในกรณีที่ข้อมูลมีความเป็นปกติ

2) กรณีที่สองเป็นการเปรียบเทียบประสิทธิภาพของตัวประมาณในกรณีที่ข้อมูลมีการปะปนด้วยข้อมูลที่เป็นค่าผิดปกติ

โดยทั้งสองกรณีจะพิจารณาตามหัวข้อต่อไปนี้

1) เมื่อขนาดตัวอย่างที่ใช้ต่างกันจะทำให้ตัวประมาณค่าต่างๆ มีประสิทธิภาพขึ้นอยู่กับขนาดตัวอย่างอย่างไร

2) สำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่น ค่า Window width ที่แตกต่างกันจะมีผลกระทบต่อประสิทธิภาพของตัวประมาณหรือไม่

3) สำหรับกรณีที่ข้อมูลมีการปะปนด้วยข้อมูลที่เป็นค่าผิดปกติ สัดส่วนค่าผิดปกติของข้อมูลที่แตกต่างกันและค่าผิดปกติที่มีระยะห่างจากข้อมูลส่วนใหญ่แตกต่างกัน จะมีผลกระทบต่อประสิทธิภาพของตัวประมาณหรือไม่

กลุ่มของการแจกแจงต่อไปนี้จะถูกใช้ในการศึกษา โดยกำหนดให้ตัวอย่างที่ศึกษามาจากประชากรที่มีการแจกแจงใน 2 ลักษณะ ดังนี้

1) การแจกแจงแบบปกติด้วยค่าเฉลี่ย μ และความแปรปรวน σ^2 หรืออาจเขียนได้เป็น $N(\mu,\sigma^2)$ ซึ่งในที่นี้กำหนดให้ $\mu = 0$ และ $\sigma^2 = 1$

2) การแจกแจงแบบ Contaminated normal หรืออาจเขียนเป็น $CN(\mu_1,\sigma^2_1, p, \mu_2,\sigma^2_2)$ เมื่อ p เป็นสัดส่วนของขนาดตัวอย่างที่สุ่มจากการแจกแจงแบบ $N(\mu_2,\sigma^2_2)$ ที่นำมาผสมกับตัวอย่างที่สุ่มจากการแจกแจงแบบ $N(\mu_1,\sigma^2_1)$ หรืออีกนัยหนึ่ง p เป็นสัดส่วนของจำนวนข้อมูลที่เป็นค่าผิดปกติ โดยในที่นี้กำหนดให้ $\mu_1 = 0$ และ $\sigma^2_1 = 1$ ส่วน $\mu_2 =$ 4, 7 และ 10 ส่วน σ^2_2 กำหนดให้เท่ากับ 1 สำหรับสัดส่วนของการผสม μ_2 และ σ^2_2 จะกำหนดให้ p = 0.03, 0.05, 0.10, 0.20 และ0.30

ขั้นตอนในการดำเนินงาน

1.1 จำลองแบบข้อมูล จากประชากรที่กำหนดคือการแจกแจงดังต่อไปนี้

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

- การแจกแจงแบบปกติด้วยค่าเฉลี่ย 0 และความแปรปรวน 1
- การแจกแจงแบบ Contaminated Normal $CN(\mu_1,\sigma^2_1,p,\mu_2,\sigma^2_2)$ โดยกำหนด $\mu_1 = 0,\sigma^2_1 = 1$ และ $\mu_2 =$ 4, 7 และ 10 โดยที่ $\sigma^2_2 = 1$ กำหนด p = 0.03, 0.05, 0.10, 0.20 และ 0.30 ตามลำดับ

ในการจำลองแบบข้อมูลจากการแจกแจงแบบ Contaminated normal ทำโดยการสร้างข้อมูล 2 ส่วน ข้อมูลส่วนแรกสุ่มมาจากการแจกแจงแบบปกติด้วยค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1 จำนวน n(1-p) ตัว ผสมกับข้อมูลส่วนที่สองซึ่งสุ่มมาจากการแจกแจงแบบปกติด้วยค่าเฉลี่ย μ_2 และความแปรปรวนเป็นσ^2_2ตามที่กำหนดไว้ข้างต้นจำนวน np ตัว(ในการคำนวณจำนวนค่าผิดปกติหากมีเศษจะทำการตัดทิ้ง)

สำหรับขนาดตัวอย่างที่ใช้กำหนดให้มีขนาดตัวอย่างเป็น 22 39 และ 100 ตามลำดับ

1.2 นำข้อมูลที่ได้จากการแจกแจงแบบต่างๆ ตามที่กำหนดมาหาค่าของตัวประมาณทั้ง 7 ตัว

ให้ $T_{n_k} = T_{n_k}(X_1, X_2, ..., X_n)$ แทนค่าประมาณที่คำนวณได้จากค่าสังเกตที่ 1 ถึง n โดยใช้ตัวประมาณที่ k

กำหนดให้

k = 1 คือ ตัวประมาณค่าเฉลี่ยตัวอย่าง

k = 2 คือ ตัวประมาณ Huber

k = 3 คือ ตัวประมาณ Huber–type skipped mean

k = 4 คือ ตัวประมาณ Three-part redescending

k = 5 คือ ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อใช้ window width เป็น $2s/n^{1/5}$

k = 6 คือ ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อใช้ window width เป็น s/2

k = 7 คือ ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อใช้ window width เป็น s/4

ตัวประมาณค่าเฉลี่ยจากค่าเฉลี่ยตัวอย่าง กำหนดโดย

$$T_{n_1} = \frac{\sum_{i=1}^{n} x_i}{n}$$

ตัวประมาณค่าเฉลี่ยจากวิธีสถิติคงทน กำหนดโดย One-step M-estimators

$$T_{n_k} = T_n^{(0)} + S_n^{(0)} \sum_{i=1}^{n} \Psi_k \left(\frac{x_i - T_n^{(0)}}{S_n^{(0)}} \right) \Big/ \sum_{i=1}^{n} \Psi'_k \left(\frac{x_i - T_n^{(0)}}{S_n^{(0)}} \right) \text{ โดยที่ k } = 2, 3, 4$$

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

เมื่อ $T_n^{(0)}$ เป็น robust estimator ของ location parameter ค่าของ $T_n^{(0)}$ ค่าแรกที่เหมาะสม กำหนดโดย

$$T_n^{(0)} = med(x_i)$$

และ $S_n^{(0)}$ เป็น robust estimator ของ scale parameter ค่าของ $S_n^{(0)}$ ค่าแรกที่เหมาะสม กำหนดโดย

$$S_n^{(0)} = 1.483 MAD(x_i) = 1.483 med_i \, | x_i - MED |$$

โดย $MED = med_i(x_i)$ คือ ค่ามัธยฐานของข้อมูล x_i โดยที่ i มีค่าตั้งแต่ 1 ถึง n และ $med_i \, | x_i - MED |$ คือ มัธยฐานของข้อมูล $| x_i - MED |$ โดยที่ i มีค่าตั้งแต่ 1 ถึง n

ในการคำนวณค่าประมาณคงทนแต่ละตัวจะใช้ค่าของ $T_n^{(0)}$ และ $S_n^{(0)}$ ที่คำนวณได้จากตัวอย่างสุ่มแต่ละชุด แทนลงใน One-step M-estimators เพื่อหาค่า T_{n_k} ซึ่งโดยปกติการหาค่า T_{n_k} จะต้องผ่านกระบวนการทำซ้ำจนกว่าค่าประมาณที่ได้ไม่แตกต่างกันมากนัก แต่ในที่นี้ใช้การแทนค่าเพียงครั้งเดียว

ตัวประมาณคงทนแต่ละตัวจะมีการกำหนด ψ-function ที่แตกต่างกันขึ้นอยู่กับเหตุผลของผู้คิดค้น

- ตัวประมาณ Huber estimator กำหนดโดย

$$\Psi_2(x) = \min\{b, \max\{x, -b\}\} = x \cdot \min\left(1, \frac{b}{|x|}\right)$$

สำหรับ $0 < b < \infty$

- ตัวประมาณ Huber-type skipped mean กำหนดโดย

$$\begin{aligned} \psi_3(x) &= x && ; \ 0 \le |x| \le r \\ &= 0 && ; \ r \le |x| \end{aligned}$$

สำหรับ $0 < r < \infty$

- ตัวประมาณ Three-part redescending M-estimator (Hampel estimator) กำหนดโดย

$$\begin{aligned} \psi_4(x) &= x && ; \ 0 \le |x| \le a \\ &= a\,sign(x) && ; \ a \le |x| \le b \\ &= a\frac{r-|x|}{r-b}sign(x) && ; \ b \le |x| \le r \\ &= 0 && ; \ r \le |x| \end{aligned}$$

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

สำหรับ $0 < a \le b < r < \infty$

สำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นมีสูตรในการคำนวณ คือ

$$T_{n_k} = \sum_{i=1}^{n} \frac{x_i \hat{f}_p(x_i)}{\sum_{i=1}^{n} \hat{f}_p(x_i)} \qquad \left[\text{เมื่อ } \sum_{i=1}^{n} \frac{\hat{f}_p(x_i)}{\sum_{i=1}^{n} \hat{f}_p(x_i)} = 1\right]$$

โดยที่ $k = 5, 6, 7$ และ $\hat{f}_p(x_i)$ คือ ค่าประมาณความหนาแน่นที่จุด x_i กำหนดโดย

$$\hat{f}_p(x_i) = \frac{1}{nh_p}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h_p}\right) \qquad \text{โดยที่ } p = 1, 2, 3$$

ซึ่ง $\hat{f}_p(\bullet)$ เป็นค่าประมาณฟังก์ชันความหนาแน่นแบบเคอร์เนล และ x_j คือค่าสังเกตที่ j เมื่อ $j = 1, 2, \ldots, n$

สำหรับ Kernel function ที่ใช้ คือ Gaussian kernel และกำหนด Window width h_p ดังนี้

h_1 มีค่าเป็น $2s/n^{1/5}$

h_2 มีค่าเป็น s/2

h_3 มีค่าเป็น s/4

1.3 ดำเนินการตามขั้นตอน 1 และ 2 ซ้ำอย่างเป็นอิสระกัน จำนวน 1,000 ครั้ง (กำหนดให้ r แทนจำนวนซ้ำ) แล้วเก็บค่าประมาณที่ได้ไว้

1.4 นำค่าประมาณที่ได้จากตัวประมาณทั้ง 7 ตัวมาหาค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานโดยใช้สูตรต่อไปนี้

ค่าเฉลี่ยของตัวประมาณค่าเฉลี่ย คือ

$$\bar{T}_{n_k} = \frac{\sum_{l=1}^{r} T_{n_{kl}}}{r}$$ เมื่อ k = 1, 2,..,7 และ l = 1, 2,…,r

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ความแปรปรวนของตัวประมาณค่าเฉลี่ย คือ

$$Var(T_{n_k}) = \frac{\sum_{l=1}^{r} (T_{n_{kl}} - \bar{T}_{n_k})}{r}$$ เมื่อ k = 1, 2,..,7 และ l = 1, 2,…,r

ความคลาดเคลื่อนมาตรฐานของตัวประมาณค่าเฉลี่ย คือ

$$S.E.(T_{n_k}) = \sqrt{\frac{\sum_{l=1}^{r} (T_{n_{kl}} - \bar{T}_{n_k})}{r}}$$ เมื่อ k = 1, 2,..,7 และ l = 1, 2,…,r

1.5 เปรียบเทียบค่าเฉลี่ยของตัวประมาณแต่ละวิธีกับค่าเฉลี่ยของประชากรจริง โดยดูว่าค่าเฉลี่ยของตัวประมาณวิธีใดใกล้เคียงกับค่าเฉลี่ยของประชากรมากกว่าแสดงว่าตัวประมาณนั้นใช้ได้ดีกว่า และพิจารณาจากความคลาดเคลื่อนมาตรฐานว่าวิธีใดให้ค่าความคลาดเคลื่อนมาตรฐานต่ำที่สุด

1.6 จากนั้นเปลี่ยนขนาดตัวอย่างที่ใช้ตามที่กำหนดไว้ในข้อ 1 ตามลำดับแล้วดำเนินการซ้ำตามขั้นตอนที่ 1–5

1.7 จากนั้นทำการจำลองแบบข้อมูลจากประชากรโดยเปลี่ยนการแจกแจงตามที่กำหนดไว้ในข้อ 1 แล้วดำเนินการซ้ำตามขั้นตอนที่ 1 – 6

2. การศึกษาที่สองเป็นการศึกษาคุณสมบัติความคงทนของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล โดยพิจารณาจาก Influence functions และ Breakdown point

เราสามารถหารูปแบบของ Influence function ของค่าเฉลี่ยจากตัวอย่าง ($\bar{X}$) และตัวประมาณคงทนอื่นๆ ได้ตามรายละเอียดในบทที่ 2 แต่เนื่องจากตัวประมาณค่าเฉลี่ยจากถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลไม่สามารถใช้นิยามของ IF ในการหารูปแบบของ IF ได้ ดังนั้นเราจึงทำการศึกษาจาก Empirical influence function (EIF) แทน โดยจะใช้โปรแกรม Fortran power station 4.0 เพื่อช่วยในการสุ่มตัวอย่าง และคำนวณค่าของ EIF

ขั้นตอนในการดำเนินงาน

2.1 จำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบปกติด้วยค่าเฉลี่ยเป็น 0 ความแปรปรวนเป็น 1 โดยกำหนดขนาดตัวอย่างเป็น 22 39 และ 100 ตามลำดับ

2.2 กำหนดค่าๆ หนึ่งในชุดข้อมูลตัวอย่างให้เป็นค่าผิดปกติ โดยเราจะแปรค่าของค่าผิดปกติ ตั้งแต่ – 500 ถึง 500 แล้วหาค่าประมาณของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อเลือก Kernel function เป็น Gaussian kernel และกำหนด Window width h เป็น $2\sigma/n^{1/5}$ และหาค่าประมาณของค่าเฉลี่ยจากตัวอย่าง

2.3 จากนั้นทำการจำลองแบบข้อมูลโดยเปลี่ยนขนาดตัวอย่างตามที่กำหนดไว้ในข้อ 1 แล้วทำซ้ำตามขั้นตอนที่ 1-2

2.4 เปรียบเทียบรูปแบบการเปลี่ยนแปลงของ EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลกับค่าเฉลี่ยตัวอย่าง

2.5 หาค่า Breakdown point ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลและค่าเฉลี่ยตัวอย่าง โดยใช้นิยามในบทที่ 2 หน้า 13

มหาวิทยาลัยเทคโนโลยีสุรนารี ลิขสิทธิ์

บทที่ 4

ผลการวิเคราะห์ข้อมูล

จากการใช้ผลลัพธ์ร่วมกันของการจำลองแบบหลายๆ แบบเพื่อศึกษาตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลตามวัตถุประสงค์ 2 ข้อ คือ

1. เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นโดยใช้วิธีการประมาณความหนาแน่นแบบเคอร์เนล (Kernel density estimation) กับตัวประมาณค่าแบบอื่นๆ คือ ค่าเฉลี่ยตัวอย่าง ($\overline{X}$) Huber estimator Huber - type skipped mean estimator และ Three - part redescending estimator

โดยแบ่งกรณีศึกษา เป็น 2 กรณี คือ

1.1 กรณีที่ไม่มีค่าผิดปกติ

1.2 กรณีที่มีค่าผิดปกติ โดยจะศึกษาทั้งในกรณีที่สัดส่วนของค่าผิดปกติมีขนาดต่างๆ กัน และกรณีที่ค่าผิดปกติมีระยะห่างจากข้อมูลส่วนใหญ่ต่างๆกัน

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

1.1 กรณีที่ไม่มีค่าผิดปกติ ทำการศึกษาตัวประมาณค่าเฉลี่ยของประชากรโดยการจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบปกติด้วยค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1 และใช้ขนาดตัวอย่าง เป็น 22 39 และ 100 ในการศึกษาจะทำการทดลองโดยการสุ่มตัวอย่างตามขนาดที่กำหนดจากประชากรจำนวน 1000 ครั้ง แล้วหาค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของตัวประมาณต่างๆ ซึ่งผลที่ได้แสดงไว้ในตารางที่ 3

จากตารางที่ 3 พบว่าเมื่อสุ่มตัวอย่างขนาด 22 ซ้ำๆ กันจากประชากรที่มีการแจกแจงแบบปกติมาตรฐานซึ่งถือว่าเป็นกรณีที่ข้อมูลไม่มีค่าผิดปกติเจือปนอยู่ ตัวประมาณทุกตัวมีค่าเฉลี่ยของค่าประมาณใกล้เคียงกันและใกล้เคียงกับค่าเฉลี่ยของประชากร โดยเฉพาะค่าเฉลี่ยจากตัวอย่างมีค่าเฉลี่ยเข้าใกล้ค่าเฉลี่ยของประชากรมากที่สุด สำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือกค่า Window width เหมาะสมจะให้ค่าเฉลี่ยของค่าประมาณใกล้เคียงกับค่าเฉลี่ยตัวอย่างและใกล้เคียงค่าเฉลี่ยของประชากรจริงมากกว่าตัวประมาณคงทนตัวอื่นๆ สำหรับการสุ่มตัวอย่างขนาด 39 และ 100 ผลที่ได้เป็นไปในทำนองเดียวกันกับกรณีตัวอย่างขนาด 22

เมื่อเพิ่มขนาดตัวอย่าง ตัวประมาณค่าเฉลี่ยทุกตัวจะมีค่าเฉลี่ยของค่าประมาณใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากขึ้น อย่างไรก็ตามค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างยังคงใกล้เคียงกว่าตัวประมาณคงทนทุกตัว ส่วนกรณีของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือก Window width เหมาะสมจะมีค่าเฉลี่ยของค่าประมาณดีเท่าหรือใกล้เคียงกับค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่าง

การเลือก Window width h สำหรับการประมาณฟังก์ชันความหนาแน่นแบบเคอร์เนลมีผลต่อการประมาณค่า เนื่องจากถ้าเลือก Window width ได้อย่างเหมาะสมค่าเฉลี่ยของค่าประมาณที่ได้จะใกล้เคียงกับค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างและค่าเฉลี่ยของประชากรจริงมากกว่าการเลือกใช้ Window width ตัวอื่นๆ

อย่างไรก็ตาม ค่าเฉลี่ยจากตัวอย่างยังคงเป็นตัวประมาณที่ดีที่สุดในกรณีที่ข้อมูลไม่มีค่าผิดปกติเจือปน เนื่องจากให้ค่าเฉลี่ยของค่าประมาณใกล้เคียงกับค่าเฉลี่ยจากประชากรมากกว่าตัวประมาณค่าเฉลี่ยตัวอื่นๆ และมีส่วนเบี่ยงเบนมาตรฐานน้อยที่สุด ซึ่งสอดคล้องกับทฤษฎีที่ว่า $\bar{X}$ เป็นตัวประมาณค่าที่ดีที่สุดและมีค่าความแปรปรวนต่ำสุดในการประมาณค่าเฉลี่ยของประชากร

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตารางที่ 3 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber – type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน (กรณีที่ไม่มีค่าผิดปกติ)

n		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
22	Mean	0.0154	0.0236	0.0196	0.0198	0.0155	0.0157	0.0164
	VAR	0.0385	0.0399	0.0388	0.0381	0.0385	0.0385	0.0384
	S.E.	0.1963	0.1997	0.1971	0.1953	0.1963	0.1961	0.1959
39	Mean	0.0085	0.0132	0.0124	0.0119	0.0085	0.0086	0.0088
	VAR	0.0254	0.0270	0.0284	0.0264	0.0254	0.0254	0.0255
	S.E.	0.1593	0.1645	0.1684	0.1626	0.1593	0.1594	0.1597
100	Mean	0.0018	0.0024	0.0019	0.0022	0.0018	0.0018	0.0019
	VAR	0.0105	0.0109	0.0109	0.0107	0.0105	0.0105	0.0105
	S.E.	0.1025	0.1045	0.1042	0.1032	0.1026	0.1026	0.1027

1.2 กรณีที่มีค่าผิดปกติ ทำการศึกษาตัวประมาณค่าเฉลี่ย โดยการจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ Contaminated normal ซึ่งกำหนดให้ข้อมูลส่วนแรกแทนข้อมูลที่เป็นปกติถูกสุ่มมาจากการแจกแจงแบบปกติด้วยค่าเฉลี่ย (μ_1) เป็น 0 และความแปรปรวน (σ_1^2) เป็น 1 จำนวน n(1-p) ตัว ผสมกับข้อมูลส่วนที่สองซึ่งเป็นข้อมูลที่เป็นค่าผิดปกติถูกสุ่มมาจากการแจกแจงแบบปกติด้วยค่าเฉลี่ย (μ_2) เป็น 4 7 และ 10 มีความแปรปรวน (σ_2^2) เป็น 1 จำนวน np ตัว กำหนดให้สัดส่วนของค่าผิดปกติ (p)เป็น 0.03 0.05 0.10 0.20 และ 0.30 และขนาดตัวอย่าง (n) เป็น 22 39 และ 100 ในการศึกษาแต่ละกรณีจะทำการทดลองโดยการสุ่มตัวอย่างตามขนาดที่กำหนดจากประชากรจำนวน 1,000 ครั้ง แล้วหาค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐาน ซึ่งผลที่ได้จะแสดงไว้ในตารางที่ 4 – 6 ต่อไปนี้

ตารางที่ 4 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 4, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 22

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.0147	0.0184	0.0154	0.0177	0.0147	0.0147	0.0149
	VAR	0.0260	0.0280	0.0274	0.0269	0.0261	0.0261	0.0262
	S.E.	0.1614	0.1673	0.1654	0.1641	0.1615	0.1616	0.1618
0.05	Mean	0.1707	0.0741	0.0328	0.0671	0.1264	0.1653	0.1562
	VAR	0.0604	0.0633	0.0903	0.0640	0.0588	0.0607	0.0601
	S.E.	0.2457	0.2515	0.3006	0.2529	0.2425	0.2463	0.2452
0.10	Mean	0.3777	0.1842	0.1129	0.1847	0.3055	0.3706	0.3559
	VAR	0.0366	0.0414	0.0604	0.0426	0.0435	0.0375	0.0394
	S.E.	0.1913	0.2035	0.2458	0.2063	0.2086	0.1935	0.1984
0.20	Mean	0.7152	0.4002	0.2995	0.4513	0.6292	0.7073	0.6904
	VAR	0.0509	0.0565	0.1186	0.0615	0.0555	0.0517	0.0535
	S.E.	0.2256	0.2378	0.3443	0.2480	0.2356	0.2273	0.2314

ตารางที่ 4 (ต่อ)

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.30	Mean	1.0917	0.8057	0.8263	0.9070	1.0194	1.0856	1.0725
	VAR	0.0427	0.0783	0.1693	0.0876	0.0537	0.0437	0.0460
	S.E.	0.2066	0.2799	0.4115	0.2959	0.2318	0.2089	0.2145

ตารางที่ 5 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 7, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 22

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.0149	0.0157	0.0170	0.0155	0.0152	0.0161	0.0174
	VAR	0.0360	0.0377	0.0654	0.0364	0.0360	0.0361	0.0364
	S.E.	0.1896	0.1942	0.2557	0.1908	0.1898	0.1901	0.1907
0.05	Mean	0.3162	0.0807	0.0411	0.0335	0.1753	0.3028	0.2760
	VAR	0.0438	0.0506	0.0917	0.0478	0.0499	0.0445	0.0461
	S.E.	0.2093	0.2249	0.3028	0.2187	0.2233	0.2110	0.2146
0.10	Mean	0.6587	0.1901	0.0629	0.1119	0.4672	0.6427	0.6095
	VAR	0.0529	0.0563	0.0718	0.0623	0.0653	0.0536	0.0555
	S.E.	0.2300	0.2373	0.2680	0.2497	0.2556	0.2315	0.2356
0.20	Mean	1.2645	0.4159	0.0693	0.3265	1.0553	1.2484	1.2141
	VAR	0.0426	0.0603	0.0850	0.0891	0.0921	0.0451	0.0535
	S.E.	0.2064	0.2456	0.2915	0.2985	0.3035	0.2124	0.2312
0.30	Mean	1.8732	0.8144	0.2682	0.8942	1.7197	1.8597	1.8319
	VAR	0.0363	0.0904	0.3142	0.1893	0.0856	0.0386	0.0475
	S.E.	0.1905	0.3007	0.5605	0.4351	0.2925	0.1964	0.2179

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตารางที่ 6 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 10, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 22

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.0159	0.0219	0.0220	0.0207	0.0161	0.0163	0.0168
	VAR	0.0429	0.0441	0.0454	0.0442	0.0429	0.0429	0.0429
	S.E.	0.2072	0.2099	0.2132	0.2103	0.2072	0.2072	0.2072
0.05	Mean	0.4471	0.0799	0.0580	0.0075	0.1997	0.4241	0.3792
	VAR	0.0484	0.0533	0.0881	0.0532	0.0628	0.0500	0.0556
	S.E.	0.2200	0.2308	0.2969	0.2307	0.2507	0.2236	0.2357
0.10	Mean	0.8830	0.1279	0.0361	0.0222	0.5805	0.8572	0.8037
	VAR	0.0375	0.0451	0.0706	0.0457	0.0752	0.0428	0.0517
	S.E.	0.1937	0.2123	0.2656	0.2137	0.2742	0.2070	0.2275
0.20	Mean	1.7992	0.4095	0.0661	0.1311	1.4821	1.7744	1.7224
	VAR	0.0401	0.0685	0.0926	0.0978	0.1411	0.0492	0.0696
	S.E.	0.2002	0.2617	0.3043	0.3128	0.3757	0.2218	0.2637
0.30	Mean	2.7347	0.7553	0.1194	0.5554	2.5091	2.7155	2.6744
	VAR	0.0464	0.0991	0.0944	0.2498	0.1302	0.0524	0.0686
	S.E.	0.2155	0.3149	0.3073	0.4998	0.3609	0.2290	0.2619

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

จากตารางที่ 4-6 เป็นผลที่ได้จากการสุ่มตัวอย่างขนาด 22 ซ้ำๆ กันจากการแจกแจงแบบ Contaminated normal ด้วย $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 4$, 7 และ 10 ตามลำดับ และ $\sigma_2^2 = 1$ ซึ่งจากผลที่ได้สามารถสรุปเป็นข้อๆ ได้ดังนี้

1) ในกรณีที่สัดส่วนของค่าผิดปกติเป็น 0.03 ซึ่งเป็นกรณีที่มีค่าผิดปกติน้อยๆ หรือไม่มีเลย ในบางตัวอย่างสุ่มตัวประมาณทุกตัวมีค่าเฉลี่ยของค่าประมาณใกล้เคียงกันและใกล้เคียงกับค่าเฉลี่ยของประชากร โดยค่าเฉลี่ยจากตัวอย่างมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยจากประชากรจริงมากที่สุดและ

มีความคลาดเคลื่อนต่ำที่สุด ค่าเฉลี่ยของค่าประมาณของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีค่าใกล้เคียงกับค่าเฉลี่ยของประชากรมากกว่าตัวประมาณคงทนต่างๆ โดยเฉพาะตัวประมาณที่ได้จากการเลือกค่า Window width ที่เหมาะสม

2) เมื่อสัดส่วนของจำนวนค่าผิดปกติเพิ่มมากขึ้นจาก 0.05 ถึง 0.30 พบว่าค่าเฉลี่ยของค่าประมาณทุกตัวมีแนวโน้มจะเบี่ยงเบนจากค่าเฉลี่ยจากประชากรจริงมากขึ้น โดยเฉพาะค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างจะแตกต่างจากค่าเฉลี่ยของประชากรจริงอย่างเห็นได้ชัด และมีความแตกต่างมากกว่าตัวประมาณค่าเฉลี่ยตัวอื่นๆ ส่วนตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือกค่า Window widthไม่เหมาะสม ค่าเฉลี่ยของค่าประมาณที่ได้จะแตกต่างจากค่าเฉลี่ยของประชากรจริงน้อยกว่าค่าเฉลี่ยจากตัวอย่างเพียงเล็กน้อย แต่สำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือกค่า Window width เหมาะสมมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากกว่าค่าเฉลี่ยจากตัวอย่างแต่แตกต่างจากค่าเฉลี่ยของประชากรมากกว่าตัวประมาณคงทนทุกตัว

3) สำหรับสัดส่วนของจำนวนค่าผิดปกติค่าหนึ่ง เมื่อขนาดของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้น ค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างและค่าเฉลี่ยของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะแตกต่างจากค่าเฉลี่ยจากประชากรจริงมากขึ้น ในขณะที่ตัวประมาณคงทนไม่เป็นเช่นนั้น คือ เมื่อขนาดของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้น ตัวประมาณ Huber จะมีค่าเฉลี่ยของค่าประมาณไม่แตกต่างกันมากนัก สำหรับตัวประมาณ Huber-type skipped mean และตัวประมาณ Hampel ค่าเฉลี่ยของค่าประมาณจะใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากขึ้น และสำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลที่เลือก Window width ที่เหมาะสม ในกรณีที่สัดส่วนของจำนวนค่าผิดปกติเป็น 0.03 ถึง 0.05 เมื่อขนาดของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้นค่าเฉลี่ยของค่าประมาณที่ได้จะไม่ค่อยแตกต่างกัน แต่ในกรณีที่สัดส่วนของจำนวนค่าผิดปกติเป็น 0.10 ถึง 0.30 เมื่อขนาดของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้นค่าเฉลี่ยของค่าประมาณที่ได้จะเบี่ยงเบนจากค่าเฉลี่ยของประชากรจริงมากขึ้นด้วยแต่จะช้ากว่าค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่าง

4) สำหรับความคลาดเคลื่อนมาตรฐานและความแปรปรวนของค่าเฉลี่ยจากตัวอย่างในแต่ละกรณีไม่แตกต่างกันมากนัก แต่สำหรับตัวประมาณอื่นๆ มีแนวโน้มจะเพิ่มขึ้นเมื่อจำนวนค่าผิดปกติเพิ่มขึ้น อย่างไรก็ตามค่าเฉลี่ยตัวอย่างยังคงมีความคลาดเคลื่อนมาตรฐานน้อยกว่าตัวประมาณตัวอื่นๆ

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

- ที่ขนาดตัวอย่างเพิ่มขึ้นเป็น 39 จากประชากรที่มีการแจกแจงแบบปกติมาตรฐานโดยมีร้อยละของการ Contaminated คือ p = 0.03 0.05 0.10 0.20 และ 0.30 โดยกำหนดการแจกแจงที่มาผสมเป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 4 7 และ 10 มีความแปรปรวนเป็น 1 ผลที่ได้จากการหาค่าเฉลี่ย ความแปรปรวนและความคลาดเคลื่อนมาตรฐานแสดงไว้ในตาราง 7 – 9 ต่อไปนี้

ตารางที่ 7 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 4, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 39

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.1173	0.0582	0.0225	0.0578	0.0800	0.1135	0.1061
	VAR	0.0283	0.0299	0.0444	0.0292	0.0313	0.0289	0.0299
	S.E.	0.1683	0.1728	0.2107	0.1709	0.1769	0.1701	0.1728
0.05	Mean	0.1189	0.0649	0.0384	0.0634	0.0846	0.1161	0.1099
	VAR	0.0201	0.0210	0.0238	0.0211	0.0208	0.0200	0.0201
	S.E.	0.1416	0.1450	0.1542	0.1454	0.1443	0.1414	0.1418
0.10	Mean	0.2987	0.1363	0.0732	0.1405	0.2229	0.2925	0.2795
	VAR	0.0221	0.0231	0.0307	0.0245	0.0244	0.0221	0.0223
	S.E.	0.1487	0.1520	0.1752	0.1566	0.1562	0.1487	0.1492
0.20	Mean	0.7180	0.4184	0.3819	0.4797	0.6199	0.7106	0.6950
	VAR	0.0285	0.0384	0.0828	0.0429	0.0401	0.0301	0.0330
	S.E.	0.1687	0.1960	0.2878	0.2070	0.2002	0.1736	0.1816
0.30	Mean	1.1276	0.8705	0.9369	0.9917	1.0433	1.1217	1.1087
	VAR	0.0229	0.0478	0.0959	0.0418	0.0301	0.0232	0.0244
	S.E.	0.1514	0.2189	0.3097	0.2045	0.1735	0.1524	0.1563

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ตารางที่ 8 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 7, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 39

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.1978	0.0552	0.0157	0.0299	0.0824	0.1859	0.1639
	VAR	0.0265	0.0261	0.0355	0.0262	0.0263	0.0266	0.0270
	S.E.	0.1627	0.1616	0.1884	0.1620	0.1622	0.1631	0.1644
0.05	Mean	0.1995	0.0603	0.0338	0.0325	0.0864	0.1876	0.1666
	VAR	0.0268	0.0293	0.0440	0.0288	0.0311	0.0282	0.0301
	S.E.	0.1638	0.1711	0.2098	0.1696	0.1762	0.1679	0.1736
0.10	Mean	0.5294	0.1358	0.0559	0.0600	0.3476	0.5138	0.4823
	VAR	0.0294	0.0338	0.0549	0.0365	0.0513	0.0334	0.0392
	S.E.	0.1716	0.1838	0.2343	0.1911	0.2265	0.1828	0.1981
0.20	Mean	1.2395	0.4100	0.0833	0.3233	1.0255	1.2233	1.1892
	VAR	0.0286	0.0429	0.0544	0.0679	0.0700	0.0330	0.0409
	S.E.	0.1692	0.2070	0.2331	0.2605	0.2645	0.1817	0.2022
0.30	Mean	1.9596	0.9289	0.3794	1.0542	1.7864	1.9475	1.9210
	VAR	0.0210	0.0718	0.3541	0.1581	0.0565	0.0234	0.0297
	S.E.	0.1450	0.2679	0.5951	0.3976	0.2377	0.1531	0.1723

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตารางที่ 9 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 10, 1) เมื่อ p = 0.03, 0.05, 0.10 0.20 และ 0.30 และตัวอย่างขนาด 39

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.2450	0.0305	0.0060	-0.0104	0.0840	0.2262	0.1921
	VAR	0.0202	0.0207	0.0374	0.0202	0.0205	0.0208	0.0213
	S.E.	0.1423	0.1439	0.1934	0.1421	0.1433	0.1441	0.1460
0.05	Mean	0.2528	0.0355	0.0134	-0.0034	0.0879	0.2330	0.1975
	VAR	0.0233	0.0265	0.0496	0.0255	0.0275	0.0239	0.0256
	S.E.	0.1526	0.1627	0.2228	0.1596	0.1658	0.1546	0.1599
0.10	Mean	0.7639	0.1354	0.0414	0.0053	0.4752	0.7378	0.6856
	VAR	0.0233	0.0293	0.0420	0.0291	0.0556	0.0280	0.0361
	S.E.	0.1526	0.1711	0.2050	0.1706	0.2358	0.1674	0.1900
0.20	Mean	1.7904	0.4158	0.0727	0.1239	1.4642	1.7661	1.7137
	VAR	0.0230	0.0310	0.0409	0.0441	0.1113	0.0295	0.0448
	S.E.	0.1518	0.1760	0.2023	0.2100	0.3335	0.1717	0.2117
0.30	Mean	2.8227	0.9391	0.0812	0.7010	2.5588	2.8046	2.7642
	VAR	0.0220	0.0519	0.0312	0.1895	0.1116	0.0260	0.0402
	S.E.	0.1484	0.2278	0.1766	0.4353	0.3341	0.1613	0.2004

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

จากตารางที่ 7-9 เป็นผลที่ได้จากการสุ่มตัวอย่างขนาด 39 ซ้ำๆกันจากการแจกแจงแบบ Contaminated Normal ด้วย $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 4$, 7 และ 10 ตามลำดับ และ $\sigma_2^2 = 1$ ซึ่งจากผลที่ได้สามารถสรุปเป็นข้อๆ ได้ดังนี้

1) เมื่อสัดส่วนของจำนวนค่าผิดปกติเพิ่มมากขึ้นจาก 0.03 ถึง 0.30 พบว่าค่าเฉลี่ยของค่าประมาณทุกตัวมีแนวโน้มจะเบี่ยงเบนจากค่าเฉลี่ยจากประชากรจริงมากขึ้น โดยเฉพาะค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างจะแตกต่างจากค่าเฉลี่ยของประชากรจริงอย่างเห็นได้ชัดและมีความ

แตกต่างมากกว่าตัวประมาณค่าเฉลี่ยตัวอื่นๆ ส่วนตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือกค่า Window width ที่เหมาะสมมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากกว่าค่าเฉลี่ยจากตัวอย่าง แต่น้อยกว่าค่าเฉลี่ยของค่าประมาณจากตัวประมาณคงทน ในขณะที่ตัวประมาณคงทนจะมีค่าเฉลี่ยของค่าประมาณแตกต่างจากค่าเฉลี่ยจากประชากรจริงน้อยกว่าเมื่อเทียบกับค่าเฉลี่ยจากตัวอย่างและตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล

2) สำหรับสัดส่วนของจำนวนค่าผิดปกติคงที่ค่าหนึ่ง เมื่อขนาดของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้น ค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างและค่าเฉลี่ยของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะแตกต่างจากค่าเฉลี่ยจากประชากรจริงมากขึ้น ในขณะที่ตัวประมาณคงทนไม่เป็นเช่นนั้น คือ เมื่อขนาดของค่าผิดปกติเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้น ตัวประมาณ Huber จะมีค่าเฉลี่ยของค่าประมาณไม่แตกต่างกันมากนัก สำหรับตัวประมาณ Huber-type skipped mean และตัวประมาณ Hampel ค่าเฉลี่ยของค่าประมาณจะแตกต่างค่าเฉลี่ยจากประชากรจริงน้อยลง และสำหรับตัวประมาณค่าเฉลี่ยจากการประมาณความหนาแน่นที่เลือก Window width ที่เหมาะสมในกรณีที่สัดส่วนของจำนวนค่าผิดปกติเป็น 0.03 ถึง 0.05 เมื่อค่าผิดปกติมีระยะห่างจากข้อมูลส่วนใหญ่มากขึ้นค่าเฉลี่ยของค่าประมาณที่ได้จะไม่ค่อยแตกต่างกัน แต่ในกรณีที่สัดส่วนของจำนวนค่าผิดปกติเป็น 0.10 ถึง 0.30 เมื่อค่าผิดปกติมีระยะห่างจากข้อมูลส่วนใหญ่มากขึ้นค่าเฉลี่ยของค่าประมาณที่ได้จะเบี่ยงเบนจากค่าเฉลี่ยของประชากรจริงมากขึ้นด้วยแต่จะช้ากว่าค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่าง

3) สำหรับความคลาดเคลื่อนมาตรฐานและความแปรปรวนของค่าเฉลี่ยจากตัวอย่างในแต่ละกรณีจะไม่แตกต่างกันมากนัก แต่สำหรับตัวประมาณอื่นๆ มีแนวโน้มจะเพิ่มขึ้นเมื่อสัดส่วนของจำนวนค่าผิดปกติเพิ่มขึ้น อย่างไรก็ตามค่าเฉลี่ยตัวอย่างยังคงมีความคลาดเคลื่อนมาตรฐานน้อยกว่าตัวประมาณตัวอื่นๆ

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

- ที่ตัวอย่างมีขนาดใหญ่โดยขนาดตัวอย่างเพิ่มขึ้นเป็น 100 จากประชากรที่มีการแจกแจงแบบปกติมาตรฐาน และกำหนดให้ข้อมูลค่าผิดปกติที่มาผสมมีการแจกแจงแบบปกติด้วยค่าเฉลี่ยเป็น 4 7 และ 10 มีความแปรปรวนเป็น 1 และมีสัดส่วนของการ Contaminated เป็น 0.03 0.05 0.10 0.20 และ 0.30 ตามลำดับ ผลที่ได้จากการหาค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานแสดงไว้ในตาราง 10 – 12 ต่อไปนี้

ตารางที่ 10 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 4, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 100

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.0815	0.0352	0.0136	0.0349	0.0445	0.0785	0.0728
	VAR	0.0102	0.0114	0.0118	0.0112	0.0111	0.0102	0.0103
	S.E.	0.1008	0.1070	0.1086	0.1057	0.1055	0.1010	0.1016
0.05	Mean	0.2019	0.0900	0.0331	0.0925	0.1271	0.1963	0.1854
	VAR	0.0102	0.0115	0.0191	0.0117	0.0126	0.0107	0.0112
	S.E.	0.1017	0.1070	0.1383	0.1079	0.1121	0.1033	0.1059
0.10	Mean	0.4044	0.1976	0.0692	0.2131	0.2780	0.3971	0.3822
	VAR	0.0094	0.0114	0.0181	0.0124	0.0156	0.0098	0.0109
	S.E.	0.0967	0.1068	0.1344	0.1111	0.1250	0.0991	0.1044
0.20	Mean	0.8035	0.4931	0.4835	0.5750	0.6764	0.7962	0.7804
	VAR	0.0099	0.0144	0.0409	0.0156	0.0224	0.0106	0.0124
	S.E.	0.0994	0.1200	0.2023	0.1248	0.1495	0.1031	0.1115
0.30	Mean	1.1997	0.9947	1.1267	1.1013	1.0983	1.1944	1.1825
	VAR	0.0095	0.0172	0.0219	0.0136	0.0211	0.0101	0.0116
	S.E.	0.0973	0.1311	0.1480	0.1166	0.1454	0.1003	0.1077

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ตารางที่ 11 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s / n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 7, 1) เมื่อ p = 0.03, 0.05, 0.10, 0.20 และ 0.30 และตัวอย่างขนาด 100

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s / n^{1/5}$	s/2	s/4
0.03	Mean	0.1410	0.0360	0.0027	0.0151	0.0439	0.1310	0.1135
	VAR	0.0096	0.0106	0.0102	0.0104	0.0108	0.0098	0.0102
	S.E.	0.0982	0.1029	0.1010	0.1022	0.1040	0.0988	0.1008
0.05	Mean	0.3500	0.0876	0.0300	0.0380	0.1645	0.3352	0.3066
	VAR	0.0088	0.0098	0.0161	0.0105	0.0147	0.0100	0.0121
	S.E.	0.0941	0.0991	0.1270	0.1022	0.1210	0.0998	0.1100
0.10	Mean	0.7019	0.1954	0.0624	0.1097	0.4143	0.6846	0.6494
	VAR	0.0098	0.0117	0.0189	0.0138	0.0308	0.0120	0.0168
	S.E.	0.0991	0.1082	0.1375	0.1177	0.1754	0.1096	0.1296
0.20	Mean	1.3998	0.4923	0.0874	0.4278	1.0816	1.3830	1.3491
	VAR	0.0089	0.0130	0.0165	0.0260	0.0662	0.0120	0.0205
	S.E.	0.0945	0.1141	0.1283	0.1613	0.2572	0.1098	0.1433
0.30	Mean	2.0971	1.0871	0.5158	1.3233	1.8789	2.0856	2.0602
	VAR	0.0098	0.0282	0.2270	0.0564	0.0558	0.0118	0.0173
	S.E.	0.0989	0.1680	0.4764	0.2375	0.2362	0.1085	0.1316

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตารางที่ 12 ค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber - type skipped mean ตัวประมาณ Hampel และตัวประมาณค่าเฉลี่ยจากค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Window width เป็น $2s/n^{1/5}$, s/2 และ s/4 โดยใช้การจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ CN(0, 1, p, 10, 1) เมื่อ p = 0.03, 0.05, 0.10 0.20 และ 0.30 และตัวอย่างขนาด 100

p		$\bar{X}$	Huber	Skipped Mean	Hampel	Based on Density Estimate		
						$2s/n^{1/5}$	s/2	s/4
0.03	Mean	0.2020	0.0350	0.0025	0.0030	0.0448	0.1841	0.1531
	VAR	0.0101	0.0106	0.0104	0.0103	0.0112	0.0110	0.0121
	S.E.	0.1003	0.1030	0.1018	0.1013	0.1057	0.1049	0.1101
0.05	Mean	0.4987	0.0877	0.0319	0.0032	0.1796	0.4742	0.4276
	VAR	0.0100	0.0114	0.0185	0.0112	0.0202	0.0121	0.0171
	S.E.	0.1000	0.1070	0.1362	0.1058	0.1421	0.1100	0.1309
0.10	Mean	0.9053	0.1762	0.0604	0.0179	0.4801	0.8785	0.8246
	VAR	0.0092	0.0104	0.0161	0.0105	0.0454	0.0133	0.0227
	S.E.	0.0959	0.1021	0.1369	0.1024	0.2131	0.1155	0.1508
0.20	Mean	1.9955	0.4961	0.0908	0.1672	1.5176	1.9717	1.9204
	VAR	0.0091	0.0138	0.0166	0.0251	0.1285	0.0155	0.0327
	S.E.	0.0952	0.1174	0.1288	0.1583	0.3584	0.1244	0.1809
0.30	Mean	3.0032	1.0838	0.0710	0.9373	2.6221	2.9858	2.9475
	VAR	0.0097	0.0276	0.0190	0.1023	0.1271	0.0157	0.0293
	S.E.	0.0985	0.1663	0.1380	0.3198	0.3565	0.1254	0.1712

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

จากตารางที่ 10-12 เป็นผลที่ได้จากการสุ่มตัวอย่างขนาด 100 ซ้ำๆกันจากการแจกแจงแบบ Contaminated Normal ด้วย $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 4$, 7 และ 10 ตามลำดับ และ $\sigma_2^2 = 1$ ซึ่งจากผลที่ได้พบว่าค่าประมาณของตัวประมาณแต่ละตัวในกรณีที่สัดส่วนของค่าผิดปกติเพิ่มขึ้น และในกรณีที่ค่าผิดปกติมีความเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้นก็จะเป็นไปในทำนองเดียวกันกับกรณีที่ตัวอย่างมีขนาด 22 และ 39 ซึ่งในกรณีที่ตัวอย่างมีขนาด 100 จะแสดงให้เห็นลักษณะหรือแนวโน้มดังกล่าวชัดเจนยิ่งขึ้น

จากผลที่ได้ในตารางที่ 4-12 สามารถสรุปโดยรวมได้เป็นข้อๆ ดังนี้

1) สำหรับข้อมูลที่มีค่าผิดปกติเจือปนอยู่ เมื่อขนาดตัวอย่างเพิ่มขึ้นค่าเฉลี่ยของค่าประมาณจากตัวประมาณค่าเฉลี่ยทุกตัวมีแนวโน้มจะเข้าใกล้ค่าเฉลี่ยจากประชากรจริง โดยตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยของ ประชากรจริงมากกว่าค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่าง แต่น้อยกว่าค่าเฉลี่ยของค่าประมาณจากตัวประมาณคงทน

2) เมื่อสัดส่วนของจำนวนค่าผิดปกติเพิ่มมากขึ้น พบว่าค่าเฉลี่ยของค่าประมาณทุกตัวมีแนวโน้มจะแตกต่างจากค่าเฉลี่ยของประชากรจริงมากขึ้น โดยเฉพาะค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างจะแตกต่างจากค่าเฉลี่ยของประชากรจริงมากกว่าตัวประมาณค่าเฉลี่ยตัวอื่นๆ ส่วนค่าเฉลี่ยของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือกค่า Window width ที่ไม่เหมาะสมจะมีลักษณะเช่นเดียวกับค่าเฉลี่ยจากตัวอย่างแต่ค่าเฉลี่ยของค่าประมาณที่ได้จะแตกต่างจากค่าเฉลี่ยของประชากรจริงน้อยกว่าค่าเฉลี่ยจากตัวอย่างเพียงเล็กน้อย และสำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือกค่า Window width ที่เหมาะสมจะมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากกว่าค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างแต่น้อยกว่าค่าเฉลี่ยของค่าประมาณจากตัวประมาณคงทน ในขณะที่ตัวประมาณคงทนจะมีค่าเฉลี่ยของค่าประมาณแตกต่างจากค่าเฉลี่ยจากประชากรจริงน้อยกว่าเมื่อเทียบกับค่าเฉลี่ยจากตัวอย่างและตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล

3) สำหรับสัดส่วนของจำนวนค่าผิดปกติคงที่ค่าหนึ่ง เมื่อค่าผิดปกติมีระยะห่างจากข้อมูลส่วนใหญ่มากขึ้น ค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างและค่าเฉลี่ยของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะแตกต่างจากค่าเฉลี่ยจากประชากรจริงมากขึ้นด้วย ในขณะที่ตัวประมาณคงทนไม่เป็นเช่นนั้น คือ เมื่อค่าผิดปกติมีระยะห่างจากข้อมูลส่วนใหญ่มากขึ้นค่าเฉลี่ยของค่าประมาณจากตัวประมาณคงทนทุกตัวจะไม่มีการเปลี่ยนแปลงมากนักและให้ค่าใกล้เคียงกับค่าเฉลี่ยของประชากรจริง สำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลเมื่อเลือก Window width ที่เหมาะสมในกรณีที่สัดส่วนของจำนวนค่าผิดปกติเป็น 0.03 ถึง 0.05 ค่าเฉลี่ยของค่าประมาณที่ได้จะไม่ค่อยแตกต่างกันเมื่อค่าผิดปกติมีความเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้น แต่ในกรณีที่สัดส่วนของจำนวนค่าผิดปกติเป็น 0.10 ถึง 0.30 ค่าเฉลี่ยของค่าประมาณที่ได้จะเบี่ยงเบนจากค่าเฉลี่ยของประชากรจริงมากขึ้นด้วยแต่จะช้ากว่าค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่าง

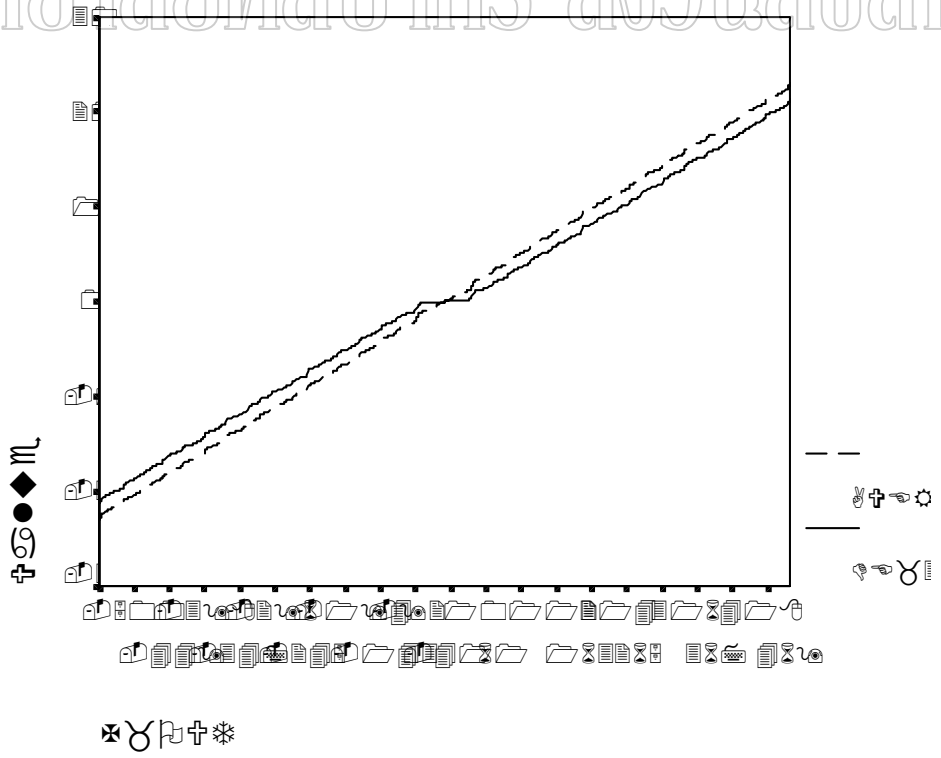
มหาวิทยาลัยศิลปากร ลิขสิทธิ์

4) สำหรับความคลาดเคลื่อนมาตรฐานและความแปรปรวนของค่าเฉลี่ยจากตัวอย่างในแต่ละกรณีจะไม่แตกต่างกันมากนัก แต่สำหรับตัวประมาณอื่นๆ มีแนวโน้มจะเพิ่มขึ้นเมื่อสัดส่วนของจำนวนค่าผิดปกติเพิ่มขึ้น อย่างไรก็ตามค่าเฉลี่ยตัวอย่างยังคงมีความคลาดเคลื่อนมาตรฐานน้อยกว่าตัวประมาณตัวอื่นๆ

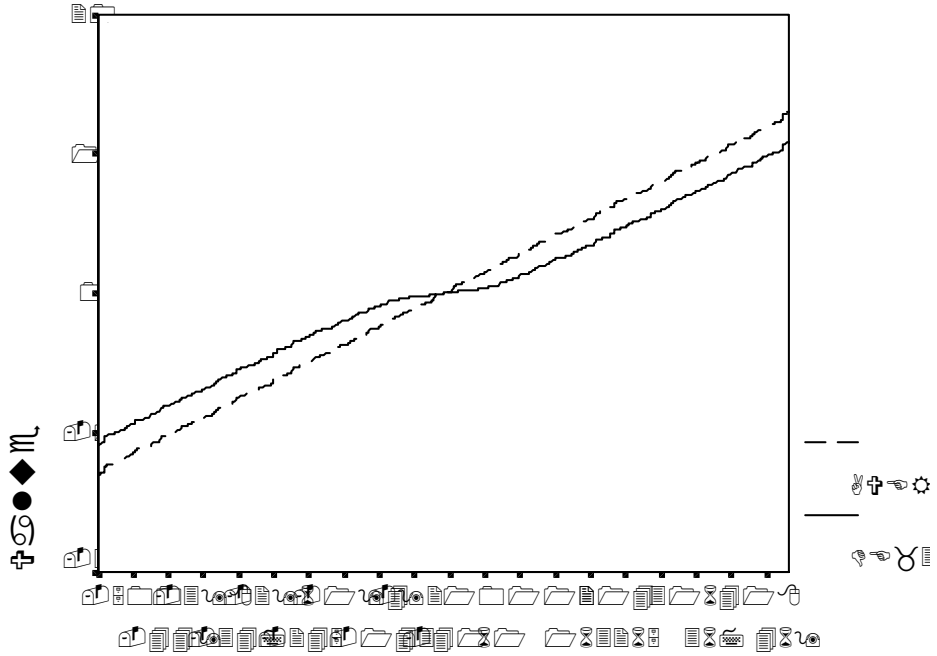
2. เพื่อศึกษาถึงคุณสมบัติความคงทนของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เราจะเลือกใช้เครื่องมือที่ใช้ในการวัดความคงทนซึ่งมีอยู่หลายวิธี เพื่อให้สามารถเปรียบเทียบความคงทนของตัวประมาณต่างๆได้ ในที่นี้จะเลือกใช้ 2 วิธีซึ่งเป็นที่รู้จักกันอย่างแพร่หลายในกลุ่มของนักสถิติคงทน คือ การใช้ Empirical influence function (EIF) และ Breakdown point (BP)

เราสามารถหารูปแบบของ Empirical influence function ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยความหนาแน่นแบบเคอร์เนลเมื่อเลือก Kernel function เป็น Gaussian kernel และ Window width เป็น $2\sigma/n^{1/5}$ เพื่อเปรียบเทียบกับ EIF ของค่าเฉลี่ยจากตัวอย่าง ($\bar{X}$) ที่ขนาดตัวอย่างเป็น 22 39 และ 100 ตามลำดับ ซึ่งผลที่ได้ถูกแสดงไว้ในภาพต่อไปนี้

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

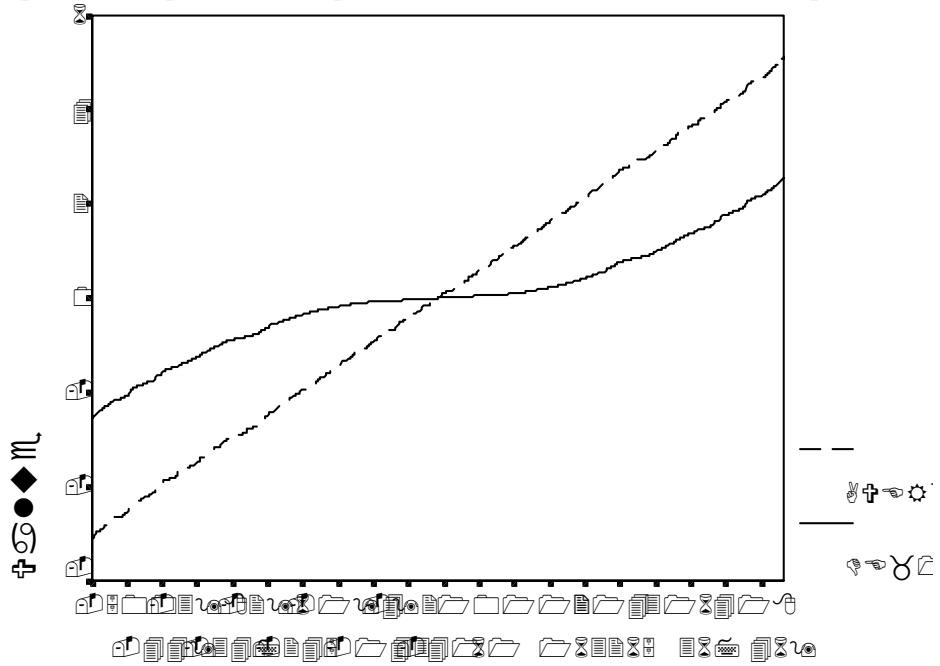


ภาพที่ 6 รูปแบบการเปลี่ยนแปลงของ EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อขนาดตัวอย่างเป็น 22



ภาพที่ 7 รูปแบบการเปลี่ยนแปลงของ EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อขนาดตัวอย่างเป็น 39

มหาวิทยาลัยศิลปากร ลิขสิทธิ์



ภาพที่ 8 รูปแบบการเปลี่ยนแปลงของ EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล เมื่อขนาดตัวอย่างเป็น 100

จากภาพที่ 6 – 8 จะเห็นได้ว่ารูปแบบ EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีลักษณะเป็นแบบค่อยๆเพิ่มขึ้นเมื่อค่าผิดปกติมีความเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้น และเมื่อขนาดตัวอย่างเพิ่มขึ้น EIF จะมีลักษณะค่อยๆ เพิ่มขึ้นอย่างช้าๆ และแม้ว่า EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะไม่มีขอบเขตที่ชัดเจนซึ่งเป็นคุณสมบัติที่ดีเหมือนกับตัวประมาณคงทน แต่ก็ไม่ได้ทำให้ค่าประมาณมีการเปลี่ยนแปลงอย่างรวดเร็วเช่นเดียวกับค่าเฉลี่ยตัวอย่างที่เมื่อค่าผิดปกติเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้นค่าประมาณก็จะเบี่ยงเบนจากค่าเฉลี่ยของประชากรจริงมากขึ้น

สำหรับเครื่องมืออีกอย่างที่ใช้ในการศึกษาคุณสมบัติความคงทนของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล ก็คือ Breakdown point โดยเราสามารถหา Breakdown point ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลและตัวประมาณค่าเฉลี่ยจากตัวอย่างตามนิยามซึ่งกล่าวไว้ในบทที่ 2 ได้ดังนี้

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

- ค่าเฉลี่ยตัวอย่าง ($\bar{X}$)

$$\varepsilon_n^* = 0\ \%$$

หมายความว่าสำหรับตัวอย่างขนาด 100 แม้จะมีค่าผิดปกติเพียงตัวเดียวก็สามารถทำให้ค่าเฉลี่ยตัวอย่างเกิดการเปลี่ยนแปลงได้

- ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณฟังก์ชันความหนาแน่นแบบเคอร์เนล

$$\varepsilon_n^* = 0\ \%$$

จะเห็นได้ว่า Breakdown Point ของตัวประมาณค่าเฉลี่ยจากค่าประมาณฟังก์ชันความหนาแน่นแบบเคอร์เนลกับค่าเฉลี่ยตัวอย่าง ($\bar{X}$) มีค่าเท่ากัน คือ 0 % นั่นคือแม้จะมีค่าผิดปกติเพียงตัวเดียวก็สามารถทำให้ค่าประมาณเกิดการเปลี่ยนแปลงได้ แต่เมื่อพิจารณาร่วมกับ EIF แล้วตัวประมาณค่าเฉลี่ยจากค่าประมาณฟังก์ชันความหนาแน่นก็ยังนับว่ามีความคงทนกว่าค่าเฉลี่ยตัวอย่างแม้จะคงทนไม่เท่ากับกลุ่มของตัวประมาณคงทนก็ตาม

บทที่ 5

สรุปผลการศึกษาและข้อเสนอแนะ

ในงานวิจัยนี้เป็นการศึกษาประสิทธิภาพของการประมาณค่าเฉลี่ยของประชากร โดยใช้ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล(Kernel density estimation) กับตัวประมาณค่าเฉลี่ยของประชากรซึ่งเป็นที่รู้จักกันอย่างแพร่หลาย คือ ค่าเฉลี่ยตัวอย่าง ($\bar{X}$) และตัวประมาณคงทนในกลุ่มของ M-estimator คือ Huber estimator Huber–type skipped mean estimator และ Three - part redescending estimator นอกจากนี้จะศึกษาคุณสมบัติของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลด้วย โดยแบ่งการศึกษาออกเป็น 2 ส่วน ในส่วนแรกเป็นการจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงตามที่กำหนดสำหรับเปรียบเทียบประสิทธิภาพของการประมาณค่าของตัวประมาณต่างๆ ข้างต้น และส่วนที่สองเป็นการศึกษาคุณสมบัติความคงทนของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล โดยจะศึกษาจาก Empirical Influence Function (EIF) ซึ่งเป็นฟังก์ชันที่ใช้อธิบายถึงอิทธิพลของการเกิดค่าผิดปกติเพียงตัวเดียวต่อค่าของตัวประมาณที่ใช้ และ Breakdown point ซึ่งเป็นเครื่องมือที่ใช้วัดขอบเขตหรือจำนวนของการเจือปนน้อยที่สุดของค่าผิดปกติที่อาจเป็นสาเหตุให้ตัวประมาณเกิดการเปลี่ยนแปลง

มหาวิทยาลัยศิลปากร ศูนย์วิทยทรัพยากร

ในส่วนแรกจะแบ่งกรณีศึกษาออกเป็น 2 กรณี คือ

1. กรณีที่ไม่มีค่าผิดปกติ ทำการศึกษาตัวประมาณค่าเฉลี่ย โดยการจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบปกติด้วยค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1

2. กรณีที่มีค่าผิดปกติ ทำการศึกษาตัวประมาณค่าเฉลี่ย โดยการจำลองแบบข้อมูลจากประชากรที่มีการแจกแจงแบบ Contaminated normal โดยกำหนดให้ข้อมูลส่วนแรกแทนข้อมูลที่เป็นปกติจากการแจกแจงแบบปกติด้วยค่าเฉลี่ย (μ_1) เป็น 0 และความแปรปรวน (σ_1^2) เป็น 1 และข้อมูลส่วนที่สองเป็นข้อมูลที่เป็นค่าผิดปกติจากการแจกแจงแบบปกติด้วยค่าเฉลี่ย (μ_2) เป็น 4 7 และ 10 มีความแปรปรวน (σ_2^2) เป็น 1 โดยมีสัดส่วนของค่าผิดปกติเป็น 0.03 0.05 0.10 0.20 และ 0.30

ทั้งสองกรณีจะใช้ขนาดตัวอย่างเป็น 22 39 และ 100 ซึ่งแทนตัวอย่างขนาดเล็ก กลาง ใหญ่ตามลำดับ โดยจะเปรียบเทียบค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของค่าประมาณที่ได้จาก

ตัวประมาณค่าเฉลี่ยซึ่งได้กล่าวไว้ข้างต้นจากการจำลองแบบข้อมูล 1,000 ครั้ง อย่างเป็นอิสระกัน

ซึ่งพบว่าในกรณีที่ไม่มีค่าผิดปกติตัวประมาณทุกตัวค่อนข้างมีความสัมพันธ์กัน คือ ตัวประมาณทุกตัวมีค่าเฉลี่ยของค่าประมาณใกล้เคียงกัน และเมื่อใช้ขนาดตัวอย่างมากขึ้นค่าเฉลี่ยของค่าประมาณของตัวประมาณค่าเฉลี่ยแต่ละตัวจะมีค่าใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากขึ้นด้วย โดยค่าเฉลี่ยจากตัวอย่างเป็นตัวประมาณที่มีประสิทธิภาพมากที่สุดเนื่องจากค่าเฉลี่ยของค่าประมาณใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากที่สุดและมีความคลาดเคลื่อนมาตรฐานต่ำที่สุด

ในกรณีที่มีค่าผิดปกติพบว่า เมื่อสัดส่วนของจำนวนค่าผิดปกติมากขึ้นค่าเฉลี่ยของตัวประมาณค่าเฉลี่ยทุกตัวจะเบี่ยงเบนจากค่าเฉลี่ยของประชากรจริงมากขึ้น และเมื่อระยะห่างของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้น ค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างจะแตกต่างจากค่าเฉลี่ยของประชากรจริงมากขึ้น ในขณะที่ตัวประมาณคงทนจะไม่ค่อยแตกต่างจากค่าเฉลี่ยของประชากรจริงมากนัก โดยในกลุ่มของตัวประมาณคงทนที่ศึกษา ตัวประมาณ Huber-type skipped mean ค่อนข้างคงทนมากกว่าตัวอื่นๆ

เนื่องจากตัวประมาณคงทนถูกถ่วงน้ำหนักด้วยฟังก์ชันที่ลดอิทธิพลของค่าผิดปกติ ดังนั้นค่าประมาณของตัวประมาณ Huber ตัวประมาณ Huber–type skipped mean และตัวประมาณ Three-part redescending จึงไม่ค่อยเปลี่ยนแปลงเมื่อระยะห่างของค่าผิดปกติแตกต่างจากข้อมูลส่วนใหญ่มากขึ้น

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

สำหรับตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล ในกรณีค่าผิดปกติเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้นแต่มีสัดส่วนของจำนวนค่าผิดปกติไม่มากนักค่าเฉลี่ยของค่าประมาณที่ได้จะค่อนข้างใกล้เคียงกับค่าเฉลี่ยของค่าเฉลี่ยจากตัวอย่างและค่าเฉลี่ยของประชากรจริง และเมื่อสัดส่วนของจำนวนค่าผิดปกติมากขึ้นค่าเฉลี่ยของค่าประมาณของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีแนวโน้มจะเบี่ยงเบนจากค่าเฉลี่ยของประชากรมากขึ้น สำหรับการเลือกค่า Window width ในการประมาณค่าความหนาแน่นจะมีผลต่อการประมาณค่า คือ ถ้าเลือกค่า Window width ได้เหมาะสมค่าประมาณที่ได้จะใกล้เคียงกับค่าเฉลี่ยของประชากรจริง

และเมื่อขนาดตัวอย่างเพิ่มขึ้นตัวประมาณค่าเฉลี่ยทุกตัวให้ค่าเฉลี่ยของค่าประมาณค่อนข้างใกล้เคียงกับค่าเฉลี่ยของประชากรจริงมากขึ้น

สำหรับในส่วนของการศึกษาเกี่ยวกับความคงทนของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล ซึ่งเราศึกษาโดยการสร้าง EIF ของตัวประมาณจากตัวอย่างขนาด 22 39 และ 100 และกำหนดให้มีค่าผิดปกติ 1 ตัวในชุดข้อมูล

ตัวอย่างซึ่งมีการเปลี่ยนแปลงค่าจาก -∞ ไปถึง +∞ พบว่ารูปแบบ EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีลักษณะเป็นแบบค่อยๆ ลดอิทธิพลของค่าผิดปกติเมื่อค่าผิดปกติมีความเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้น และแม้ว่า EIF ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจะไม่มีขอบเขตที่ชัดเจนซึ่งเป็นคุณสมบัติที่ดีเหมือนกับตัวประมาณคงทน แต่ก็ไม่ได้ทำให้ค่าประมาณมีการเปลี่ยนแปลงอย่างรวดเร็วเช่นเดียวกับค่าเฉลี่ยจากตัวอย่างเมื่อค่าผิดปกติเบี่ยงเบนจากข้อมูลส่วนใหญ่มากขึ้น สำหรับ Breakdown Point ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลมีค่าเท่ากันกับ Breakdown Point ของค่าเฉลี่ยจากตัวอย่าง ($\bar{X}$) นั่นคือ แม้จะมีค่าผิดปกติเพียงตัวเดียวก็สามารถทำให้ค่าประมาณเกิดการเปลี่ยนแปลงได้ แต่เมื่อพิจารณาร่วมกับ EIF แล้วตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลก็ยังนับว่ามีความคงทนมากกว่าค่าเฉลี่ยตัวอย่างแม้จะมีความคงทนไม่เท่ากับตัวประมาณในกลุ่มของตัวประมาณคงทนก็ตาม

ดังนั้นตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลจึงนับเป็นทางเลือกหนึ่งที่น่าสนใจ หากข้อมูลที่ต้องการศึกษาเป็นข้อมูลที่ไม่ทราบการแจกแจงและผู้วิจัยไม่แน่ใจว่าจะมีค่าผิดปกติหรือไม่ เนื่องจากในกรณีที่ไม่มีค่าผิดปกติตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นมีค่าประมาณค่อนข้างใกล้เคียงกับค่าเฉลี่ยจากตัวอย่าง และกรณีที่มีจำนวนค่าผิดปกติไม่มากนักตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นจะค่อนข้างคงทนกว่าค่าเฉลี่ยตัวอย่างถึงแม้ว่าจะมีความคงทนไม่เท่ากับตัวประมาณคงทนก็ตาม

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ข้อเสนอแนะเพื่อการวิจัยค้นคว้าครั้งต่อไป

การขยายการศึกษาสำหรับการประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลซึ่งอาจมีประโยชน์ในทางปฏิบัติ เช่น ศึกษาตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นในกรณีที่ข้อมูลมีการแจกแจงแบบอื่นๆ หรืออาจศึกษา Window width หรือ Kernel function ที่มีความเหมาะสมกับการแจกแจงใดการแจกแจงหนึ่งซึ่งจะส่งผลให้ค่าประมาณที่ได้มีคุณสมบัติที่ดีขึ้น หรือศึกษาประสิทธิภาพของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลที่ใช้ Kernel function ที่มีลักษณะของปลายหางยาวๆ ซึ่งจะให้น้ำหนักน้อยลงแก่ค่าสังเกตที่เป็นค่าผิดปกติ

บรรณานุกรม

Allen, D. L. “Hypothesis Testing Using an L1-Distance Bootstrap.” The American Statistician 51(1997) : 145-150.

Bernoulli, D. “Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda.” Acta Acad. Sci. Petropolit 1(1777) : 3-33.

Bessel, F. W. Fundamenta Astronomiae. Nicolovius : Konigsberg, 1818.

Bessel, F. W. ,and J. J. Baeyer. “Gradmessung in Ostpreussen und ihre Verbindung mit Preussischen und Russischen Dreiecksketten.” Druckerei der Koniglichen Akademie der Wissenschaften Berlin 3(1876) : 62-138.

Boscovich, R. J. “De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plurab eius ex exemplaria etiam sensorum impressa.” Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii 4(1757) : 353-396.

Chauvenet, W. “Method of Least Squares.” Manual of Spherical and Practical Astronomy 2(1863) : 469-566.

Cushny, A. R. ,and A. R. Peebles. “The action of optical isomers II.” Hyoscines. J . Physiol 32(1905) : 501-510.

Devroye, L. , and L. Gyorfi. Nonparametric Density Estimation the L1 View. New York : Wiley, 1985.

Edgeworth, F. Y. “The method of least squares.” Phios. Mag 23(1883) : 364-375.

Frees, E. W. “Estimating Densities of Functions of Observation.” Jornal of the American Statistical Association 89(1994) : 517 – 525.

Hampel, F. R. “Contributions to the theory of robust estimation.” Ph.D. thesis, University of California, Berkerey , 1968.

Hampel, F. R. “Robust Statistics : A brief introduction and overview.” Seminar for Statistics, ETH Zurich, Switzerland , 2001.

Hampel, F. R. et al. The Approach Based on Influence Functions. New York : John Wiley & Sons, Inc., 1986.

Hoeffding, W. “A Class of Statistics With Asymptotically Normal Distribution.” Annals of Mathematical Statistics 19(1948) : 193-225.

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

Huber, P. J. “Robust Estimation of a Location Parameter.” Annals of Mathematical Statistics 35(1964) : 73-101.

Huber, P. J. “Robust Statistics : A review.” Annals of Mathematical Statistics 43(1972) : 1041-1067.

Huber, P. J. Robust Statistics. New York : John Wiley & Sons, Inc., 1981.

Izenman, A. J. “ Recent Developmenta in Nonparametric Density Estimation.” Jornal of the American Statistical Association 86(1991) : 205-224.

Jeffrey, H. Theory of Probability. Clarendon Press, Oxford, 1939, 1948, 1961.

Mendeleev, D. I. “Course of work on the renewal of prototypes or standard measures of lengths and weighs (Russian).” Vremennik Glavnoi Palaty Mer I Vesov 2(1895) : 157-185.

Newcomb, S. “A generalized theory of the combination of observations so as to obtain the best result.” American Journal Mathematics 8(1886) : 343-366.

Peirce, B. “Criterion for the rejection of doubtful observations.” Astr. J 2(1852) : 161-163.

Silverman, B. W. Density Estimation For Statistics and Data Analysis. London : Chapman and Hall ,1986.

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

Student “Error of routine analysis.” Biometrika 19(1927) : 151-164.

Tukey, J. W. “A survey of sampling from contaminated distributions.” In Contributions to Probability and Statistics I. Olkin(ed.). Stanford University Press, Stanford, Calif. (1960) : 448-485.

Wright, T. W. A Treatise on the Adjustment of Observations by the Method of Least Squares. New York : Van Nostrand, 1884.

ภาคผนวก ก

โปรแกรมสำหรับคำนวณค่าสถิติที่ใช้ในงานวิจัย

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

โปรแกรมสำหรับคำนวณค่าเฉลี่ย ความแปรปรวน และความคลาดเคลื่อนมาตรฐานของตัวประมาณค่าเฉลี่ยทั้ง 7 ตัว ได้แก่ ตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นที่เลือกค่า window width เป็น h s/2 และ s/4 ค่าเฉลี่ยจากตัวอย่าง ตัวประมาณ Huber ตัวประมาณ Huber-type skipped mean และตัวประมาณ Three-part redecsending

1. ความหมายของตัวแปรในโปรแกรม

ตัวแปรในส่วนของโปรแกรมหลัก มีดังนี้

n	แทน ขนาดตัวอย่าง
r	แทน จำนวนของชุดตัวอย่างที่สุ่มซ้ำ ในกรณีเดียวกัน
tn(1000,7)	แทน ค่าประมาณของตัวประมาณแต่ละตัว ซึ่งมี 7 ตัว ซึ่งคำนวณสำหรับ ตัวอย่างแต่ละชุด ทั้งหมด 1,000 ชุด ดังนั้นแต่ละตัวประมาณจะมีค่าประมาณ 1,000 ค่าสำหรับตัวอย่างแต่ละชุด
summ(7)	แทน ผลรวมของค่าประมาณของตัวประมาณแต่ละตัว ซึ่งมี 7 ตัว
xbar(7)	แทน ค่าเฉลี่ยของค่าประมาณของตัวประมาณแต่ละตัว ซึ่งมี 7 ตัว
stdv(7)	แทน ความคลาดเคลื่อนมาตรฐานของค่าประมาณของตัวประมาณแต่ละตัว ซึ่งมี 7 ตัว
vari(7)	แทน ความแปรปรวนของค่าประมาณของตัวประมาณแต่ละตัวซึ่งมี 7 ตัว
rn(22)	แทน ตัวอย่างสุ่มที่มีขนาดเท่าที่กำหนด คือ 22 39 และ100 เช่น ในที่นี้ให้มีขนาดตัวอย่างเป็น 22
mn	แทน ค่าเฉลี่ยตัวอย่าง
med	แทน ค่ามัธยฐาน
sn	แทน Scale parameter ของตัวประมาณคงทน
xts(22)	แทน ค่าของตัวอย่างสุ่มที่ถูกแปลงสำหรับใช้ประมาณค่าสำหรับตัวประมาณคงทน
me	แทน ค่าประมาณจากตัวประมาณคงทน
den	แทน ค่าประมาณของตัวประมาณค่าเฉลี่ยจากการประมาณฟังก์ชันความหนาแน่น
tmp	แทน ผลรวมกำลังสองของส่วนเบี่ยงเบนของค่าสังเกตจากค่าเฉลี่ยของตัวประมาณ

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

avr แทน ค่าเฉลี่ยของประชากรจากการแจกแจงแบบปกติ

avr1 แทน ค่าเฉลี่ยของข้อมูลส่วนที่ 1 จากการแจกแจงแบบ Contaminated normal

avr2 แทน ค่าเฉลี่ยของข้อมูลส่วนที่ 2 จากการแจกแจงแบบ Contaminated normal

var แทน ความแปรปรวนของประชากรจากการแจกแจงแบบปกติ

var1 แทน ความแปรปรวนของข้อมูลส่วนที่ 1 จากการแจกแจงแบบ Contaminated normal

var2 แทน ความแปรปรวนของข้อมูลส่วนที่ 2 จากการแจกแจงแบบ Contaminated normal

p แทน สัดส่วนของการเจือปนของข้อมูลส่วนที่ 2 เทียบกับ ข้อมูลส่วนที่ 1 สำหรับการแจกแจงแบบ Contaminated normal

z แทน ตัวแปรที่กำหนดการแจกแจงของตัวอย่างสุ่ม
โดยถ้า z = 1 ตัวอย่างจะถูกสุ่มมาจากการแจกแจงแบบปกติ
และถ้า z = 2 ตัวอย่างจะถูกสุ่มมาจากการแจกแจงแบบ Contaminated normal

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตัวแปรในส่วนของ subroutine rand1(avr,var,xobs,sz) มีดังนี้

sz แทน ขนาดตัวอย่าง ตามที่กำหนดซึ่งก็คือ 22 39 และ 100

x(sz) แทน ตัวอย่างสุ่มที่สุ่มมาจากการแจกแจงแบบเอกรูป

xobs(sz) แทน ตัวอย่างสุ่มที่สุ่มมาจากการแจกแจงแบบปกติ

avr แทน ค่าเฉลี่ยของประชากร

var แทน ความแปรปรวนของประชากร

zeed แทน ค่าใดๆ ที่สุ่มมาจากการแจกแจงแบบเอกรูป มีค่าระหว่าง 0 ถึง 1

iseed แทน ค่าเริ่มต้นสำหรับการสุ่มตัวอย่าง มีค่าระหว่าง 0 ถึง 2147483646

ตัวแปรในส่วนของ subroutine rand2(avr1,avr2,var1,var2,p,xobs,sz) มีดังนี้

n1 แทน ขนาดตัวอย่างที่สุ่มจากการแจกแจง N(avr2, var2)

n2 แทน ขนาดตัวอย่างที่สุ่มจากการแจกแจง N(avr1, var2)

sz แทน ขนาดตัวอย่าง ตามที่กำหนดซึ่งก็คือ 22 39 และ 100

x(sz) แทน ตัวอย่างสุ่มที่สุ่มมาจากการแจกแจงแบบเอกรูป

xobs(sz) แทน ตัวอย่างสุ่มที่สุ่มมาจากการแจกแจงแบบ Contaminated normal : CN(avr1, var1, p, avr2, var 2)

avr1 แทน ค่าเฉลี่ยของข้อมูลส่วนที่ 1 สำหรับแจกแจงแบบ Contaminated normal

avr2 แทน ค่าเฉลี่ยของข้อมูลส่วนที่ 2 สำหรับแจกแจงแบบ Contaminated normal

var1 แทน ความแปรปรวนของข้อมูลส่วนที่ 1 แจกแจงแบบ Contaminated normal

var2 แทน ความแปรปรวนของข้อมูลส่วนที่ 2 แจกแจงแบบ Contaminated normal

zeed แทน ค่าใดๆ ที่สุ่มมาจากการแจกแจงแบบเอกรูป มีค่าระหว่าง 0 ถึง 1

iseed แทน ค่าเริ่มต้นสำหรับการสุ่มตัวอย่าง มีค่าระหว่าง 0 ถึง 2147483646

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ตัวแปรในส่วนของ subroutine bbsort(x,n) มีดังนี้

n แทน จำนวนข้อมูล

x(n) แทน ค่าของข้อมูลจำนวน n ตัว

s แทน ตัวแปรสำหรับการเปลี่ยนค่า

ตัวแปรในส่วนของ subroutine mean(mn,x,n) มีดังนี้

n แทน ขนาดของตัวอย่างสุ่ม

x(n) แทน ตัวอย่างสุ่มขนาด n

mn แทน ค่าเฉลี่ยตัวอย่าง

sum แทน ผลรวมของค่าจากตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine median(md,x,n) มีดังนี้

n แทน ขนาดของตัวอย่างสุ่ม

pos แทน ตำแหน่งกลางของข้อมูล

x(n) แทน ตัวอย่างสุ่มขนาด n

md แทน ค่ามัธยฐานของชุดตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine std(sd,x,n) มีดังนี้

n	แทน ขนาดของตัวอย่างสุ่ม
x(n)	แทน ตัวอย่างสุ่มขนาด n
sd	แทน ส่วนเบี่ยงเบนมาตรฐานของตัวอย่างสุ่ม
tmpmn	แทน ค่าเฉลี่ยของตัวอย่างสุ่ม
sum	แทน ผลรวมของค่าจากตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine rob_est(sn,x,n) มีดังนี้

n	แทน ขนาดของตัวอย่างสุ่ม
x(n)	แทน ตัวอย่างสุ่มขนาด n
med	แทน ค่ามัยฐานของข้อมูล x(n)
sn	แทน Scale parameter สำหรับตัวประมาณคงทน
tmp(n)	แทน ค่าสัมบูรณ์ของส่วนเบี่ยงเบนของค่าของตัวอย่างสุ่มจากค่ามัธยฐาน
tmpmed	แทน ค่ามัยฐานของข้อมูล tmp(n)

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ตัวแปรในส่วนของ subroutine transform_xts(xts,x,med,sn,n) มีดังนี้

n	แทน ขนาดของตัวอย่างสุ่ม
x(n)	แทน ตัวอย่างสุ่มขนาด n
xts(n)	แทน ค่าที่แปลงจากค่าของตัวอย่างสุ่มขนาด n สำหรับคำนวณตัวประมาณคงทน
med	แทน ค่ามัยฐานของข้อมูล x(n)
sn	แทน Scale parameter สำหรับตัวประมาณคงทน

ตัวแปรในส่วนของ subroutine huber_est(hub,x) มีดังนี้

hub	แทน ค่าของฟังก์ชัน ψ_b
x	แทน ค่าที่แปลงจากค่าของตัวอย่างสุ่มสำหรับคำนวณตัวประมาณคงทน
b	แทน ค่าคงที่สำหรับตัวประมาณ Huber มีค่าเป็น 1.339

ตัวแปรในส่วนของ subroutine diff_huber_est(hub,x) มีดังนี้

hub	แทน	ค่าของอนุพันธ์ของฟังก์ชัน ψ_b
x	แทน	ค่าที่แปลงจากค่าของตัวอย่างสุ่มสำหรับคำนวณตัวประมาณคงทน
b	แทน	ค่าคงที่สำหรับตัวประมาณ Huber มีค่าเป็น 1.339

ตัวแปรในส่วนของ subroutine huber_skip(hub,x) มีดังนี้

hub	แทน	ค่าของฟังก์ชัน $\psi_{SK(r)}$
x	แทน	ค่าที่แปลงจากค่าของตัวอย่างสุ่มสำหรับคำนวณตัวประมาณคงทน
r	แทน	ค่าคงที่สำหรับตัวประมาณ Huber-type skipped mean มีค่าเป็น 1.339

ตัวแปรในส่วนของ subroutine diff_huber_skip(hub,x) มีดังนี้

hub	แทน	ค่าของอนุพันธ์ของฟังก์ชัน $\psi_{SK(r)}$
x	แทน	ค่าที่แปลงจากค่าของตัวอย่างสุ่มสำหรับคำนวณตัวประมาณคงทน
r	แทน	ค่าคงที่สำหรับตัวประมาณ Huber-type skipped mean มีค่าเป็น 1.339

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

ตัวแปรในส่วนของ subroutine three_part(hub,x) มีดังนี้

hub	แทน	ค่าของฟังก์ชัน $\psi_{a,b,r}$
x	แทน	ค่าที่แปลงจากค่าของตัวอย่างสุ่มสำหรับคำนวณตัวประมาณคงทน
a , b , r	แทน	ค่าคงที่สำหรับตัวประมาณ Three part redesending มีค่าเป็น 1.7 3.4 และ 8.5 ตามลำดับ
tmpx	แทน	ค่าสัมบูรณ์ของค่าที่แปลงจากค่าของตัวอย่างสุ่ม
sig	แทน	เครื่องหมายของค่าของข้อมูล

ตัวแปรในส่วนของ subroutine diff_three_part(hub,x) มีดังนี้

hub	แทน	ค่าของอนุพันธ์ของฟังก์ชัน $\psi_{a,b,r}$
x	แทน	ค่าที่แปลงจากค่าของตัวอย่างสุ่มสำหรับคำนวณตัวประมาณคงทน
a , b , r	แทน	ค่าคงที่สำหรับตัวประมาณ Three part redesending มีค่าเป็น 1.7 3.4 และ 8.5 ตามลำดับ

tmpx แทน ค่าสัมบูรณ์ของค่าที่แปลงจากค่าของตัวอย่างสุ่ม

sig แทน เครื่องหมายของค่าของข้อมูล

ตัวแปรในส่วนของ subroutine m_est(me,med,x,xts,sn,n,sel) มีดังนี้

n แทน ขนาดตัวอย่าง

sel แทน ตัวแปรที่ใช้เลือกตัวประมาณคงทน

โดย sel = 1 ถ้าต้องการประมาณค่าของตัวประมาณ Huber

sel = 2 ถ้าต้องการประมาณค่าของตัวประมาณ Huber-type skipped mean

sel = 3 ถ้าต้องการประมาณค่าของตัวประมาณ Three part redesending

x(n) แทน ค่าของตัวอย่างสุ่ม

xts(n) แทน ค่าที่แปลงมาจากค่าของตัวอย่างสุ่ม

hub แทน ค่าของฟังก์ชัน ψ ณ ค่าสังเกตต่างๆ

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

sumhub แทน ผลรวมของค่าของฟังก์ชัน ψ ณ ค่าสังเกตต่างๆ

diff แทน ค่าของอนุพันธ์ของฟังก์ชัน ψ ณ ค่าสังเกตต่างๆ

sumdiff แทน ผลรวมของค่าของอนุพันธ์ของฟังก์ชัน ψ ณ ค่าสังเกตต่างๆ

sn แทน Scale parameter สำหรับตัวประมาณคงทน

me แทน ค่าประมาณของตัวประมาณคงทน

med แทน ค่ามัธยฐานของชุดข้อมูล

ตัวแปรในส่วนของ subroutine norm_fun(nor,x) มีดังนี้

n แทน ขนาดตัวอย่าง

nor แทน ค่าของฟังก์ชันความหนาแน่นความน่าจะเป็น ณ จุด x

x แทน ค่าของตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine Tn_h(tn,x,n,sel) มีดังนี้

n แทน ขนาดตัวอย่าง

nor แทน ค่าของ Kernel function ณ จุด x

x(n) แทน ค่าของตัวอย่างสุ่มขนาด n

sel	แทน ตัวแปรที่ใช้เลือกค่าของ Window width
	โดย sel = 1 ถ้าต้องการเลือก Window width เป็น $2s/n^{1/5}$
	sel = 2 ถ้าต้องการเลือก Window width เป็น (1/2)s
	sel = 3 ถ้าต้องการเลือก Window width เป็น (1/4)s
tn	แทน ค่าประมาณของคัวประมาณค่าเฉลี่ยจากค่าประมาณฟังก์ชันความหนาแน่น
sd	แทน ส่วนเบี่ยงเบนมาตรฐานของค่าของตัวอย่างสุ่ม
sumfx	แทน ผลรวมของค่า Kernel function
h	แทน ค่าของ Window width
tmp	แทน ค่าที่แปลงจากค่าของตัวอย่างสุ่ม
tmp1	แทน ผลคูณของค่าสังเกต x กับความน่าจะเป็นที่ได้จากการประมาณฟังก์ชันความหนาแน่นความน่าจะเป็นแบบ Kernel ณ จุด x
sum1	แทน ผลรวมของค่า Kernel function
sum2	แทน ผลรวมของผลคูณของค่าสังเกต x กับความน่าจะเป็นที่ได้จากการประมาณฟังก์ชันความหนาแน่นความน่าจะเป็นแบบ Kernel ณ จุด x
fx(n)	แทน ค่าของฟังก์ชันประมาณความหนาแน่นความน่าจะเป็นแบบ Kernel ณ จุด x

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

2. ขั้นตอนการทำงาน

2.1 จำลองแบบข้อมูลจากการแจกแจงและขนาดตัวอย่างตามที่กำหนดในบทที่ 3

2.2 ประมาณค่าของตัวประมาณค่าเฉลี่ยแต่ละตัว

2.3 ทำซ้ำจำนวน 1,000 ครั้ง

2.4 คำนวณค่าเฉลี่ย พร้อมทั้งค่าความแปรปรวน และความคลาดเคลื่อนมาตรฐานของค่าประมาณจากตัวประมาณค่าเฉลี่ยแต่ละตัว

3. แสดงรายละเอียดของโปรแกรม

```
!program comparison statistics
!implicit none
use  msimsl

integer n,r
real(4) tn(1000,7),summ(7),xbar(7),stdv(7),vari(7)
real(4) rn(22),mn,med,sn,xts(22),me,den
real(4) tmp,avr,avr1,avr2,var,var1,var2,p
integer i,j,z

n=22
r=1000
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
OPEN (3,FILE='.\output.txt')
do i=1,7
        summ(i)=0
end do
! input mean and variance for random sample
print*,'press 1 for normal or press 2 for contaminate normal'
read*,z
        if(z .eq. 1) then
                print*,'with mean of gr.1'
                read*,avr
                print*,'variance of gr.1'
                read*,var
        elseif (z .eq. 2) then
                print*,'with mean of gr.1'
```

```
        read*,avr1
        print*,'variance of gr.1'
        read*,var1
        print*,'with mean of gr.2'
        read*,avr2
        print*,'variance of gr.2'
        read*,var2
        print*,'with prob.'
        read*,p
else
        print*,'you can press 1 or 2 only!'
endif
do i=1,r
        ! random sample
        if(z .eq. 1) then
                call rand1(avr,var,rn,size(rn))
        elseif (z .eq. 2) then
                call rand2(avr1,avr2,var1,var2,p,rn,size(rn))
        endif

        call bbsort(rn,size(rn))
        !calculate Mean(Tn1)
        call mean(mn,rn,size(rn))
        print *,'mean = ',mn
        tn(i,1) = mn
        summ(1) = summ(1) + mn

        call median(med,rn,size(rn))
        call rob_est(sn,rn,size(rn))
        call transform_xts(xts,rn,med,sn,size(rn))
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
!calculate Huber est.(Tn2)
call m_est(me,med,rn,xts,sn,size(rn),1)
print *,'Tn (huber est.) = ',me
tn(i,2) = me
summ(2) = summ(2) + me

!calculate Huber skipped(Tn3)
me=0
call m_est(me,med,rn,xts,sn,size(rn),2)
print *,'Tn (huber skip.) = ',me
tn(i,3) = me
summ(3) = summ(3) + me

!calculate Three part(Tn4)
me=0
call m_est(me,med,rn,xts,sn,size(rn),3)
print *,'Tn (Three part) = ',me
tn(i,4) = me
summ(4) = summ(4) + me


!calculate den. by h(Tn5)
call Tn_h(den,rn,size(rn),1)
print *,'Tn (density by h) = ',den
tn(i,5) = den
summ(5) = summ(5) + den

!calculate den. by 2h(Tn6)
call Tn_h(den,rn,size(rn),2)
print *,'Tn (density by 2h) = ',den
```

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

```
            tn(i,6) = den
            summ(6) = summ(6) + den

            !calculate den. by h/2(Tn7)
            call Tn_h(den,rn,size(rn),3)
            print *,'Tn (density by h/2) = ',den
            tn(i,7) = den
            summ(7) = summ(7) + den

            print *,'======================'

        write(3,1)tn(i,1),tn(i,2),tn(i,3),tn(i,4),tn(i,5),tn(i,6),tn(i,7)
        end do
        write(3,*)' '
    write(3,*)'====================Summary======================='
        write(3,*)'       XBar','    Variance','       Std.'
        do i=1,7
            xbar(i) = summ(i)/r
            tmp=0
            do j=1,r
                tmp = tmp+((tn(j,i)-xbar(i))**2)
            end do
            vari(i) = tmp/(r)
            stdv(i) = sqrt(vari(i))
            print *,'no.: ',i
            print *,'Xbar = ',xbar(i)
            print *,'Var = ',vari(i)
            print *,'Std. = ',stdv(i)
```

มหาวิทยาลัยศิลปากร ลงวนลิขสิทธิ์

```
                write(3,2)xbar(i),vari(i),stdv(i)
            end do

  1     format(F10.4,F10.4,F10.4,F10.4,F10.4,F10.4,F10.4)
  2     format(F12.4,F12.4,F12.4)
            read '(F10.2)',tmp
            end

!****************************************
!
!Generate random normal distribution number ;normal
!
!****************************************
        subroutine rand1(avr,var,xobs,sz)
            implicit none
        integer i,sz
        real(4) x(sz),xobs(sz)
        real(4) avr,var,zeed,iseed

            call random_seed()
            call random_number(zeed)
            iseed = zeed * 2147483646
            CALL RNSET (iseed)
            CALL RNNOR (sz, x)
              do 5 i=1,sz
              xobs(i) = (x(i)*(var**0.5)) + avr
        5     continue
            end
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
!****************************************
!
!Generate random normal distribution number ;contaminate normal
!
!****************************************
                subroutine rand2(avr1,avr2,var1,var2,p,xobs,sz)
                        implicit none
                integer i,n1,n2,sz
                real(4) x(sz),xobs(sz)
                real(4) avr1,avr2,var1,var2,p,zeed,iseed

                call random_seed()
                call random_number(zeed)
                iseed = zeed * 2147483646
                CALL RNSET (iseed)
                CALL RNNOR (sz, x)
                 n1 = sz*p
                 n2 = sz-n1
                 do 15 i=1,n2
                        xobs(i) = (x(i)*(var1**0.5)) + avr1
        15      continue
                do 25 i=(n2+1),sz
                        xobs(i) = (x(i)*(var2**0.5)) + avr2
        25      continue
                end
```

มหาวิทยาลัยศิลปากร ลงวนลิขสิทธิ์

```
!****************************************
!
!Bubble sort (ascending order)
!
!****************************************
      subroutine bbsort(x,n)
            implicit none
      integer n,i,j
      real(4) x(n),s

      do 50 i=1,n-1
            do 50 j=1,n-i
                  if(x(j).GT.x(j+1)) then
                        s=x(j)
                        x(j)=x(j+1)
                        x(j+1)=s
                  end if
50    continue
      return
      end

!****************************************
!
!Calculate mean
!
!****************************************
      subroutine mean(mn,x,n)
            implicit none
      integer n,i
      real(4) x(n),mn,sum
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
        sum=0
        do i=1,size(x)
                sum = sum+x(i)
        end do
        mn = sum/size(x)
        return
        end

!****************************************
!
!Calculate median Tn(0)
!
!****************************************
        subroutine median(md,x,n)
                implicit none
        integer n,pos
        real(4) x(n),md

        pos=n/2
        if(mod(n,2)==0) then
                md = (x(pos)+x(pos+1))/2
        else
                md = x(pos+1)
        end if
        return
        end
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
!****************************************
!
!Calculate standardivation
!
!****************************************
        subroutine std(sd,x,n)
                implicit none
        integer n,i
        real(4) x(n),sd,tmpmn
        real(4) sum

        call mean(tmpmn,x,n)
        do i=1,size(x)
                sum=sum+((x(i)-tmpmn)**2)
        end do
        sd = sqrt(sum/(size(x)-1))
        return
        end

!****************************************
!
!Calculate Sn
!
!****************************************
        subroutine rob_est(sn,x,n)
                implicit none
        integer n,i
        real(4) x(n),med,sn
        real(4) tmpmed,tmp(n)
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
        call median(med,x,n)
        do i=1,size(x)
                tmp(i)=abs(x(i)-med)
        end do
        call bbsort(tmp,size(tmp))
        call median(tmpmed,tmp,n)
        sn = 1.483*tmpmed
        return
        end

!****************************************
!
!Transform to (Xi-Tn(0))/Sn(0)
!
!****************************************
        subroutine transform_xts(xts,x,med,sn,n)
                implicit none
        integer n,i
        real(4) x(n),xts(n)
        real(4) med,sn

        do i=1,size(x)
                xts(i)=(x(i)-med)/sn
        end do
        return
        end
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
!****************************************
!
!Calculate Huber estimator
!
!****************************************
      subroutine huber_est(hub,x)
            implicit none
      real(4) hub,x,b

      b=1.339
      if(x>=b) then
            hub = b
      elseif(x<=-b) then
            hub = -b
      else
            hub = x
      end if
      return
      end

!****************************************
!
!Calculate Differentiate of Huber estimator
!
!****************************************
      subroutine diff_huber_est(hub,x)
            implicit none
      real(4) hub,x,b
      b=1.339
      if(abs(x)>=b) then
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
            hub = 0
        else
            hub = 1
        end if
        return
        end

!****************************************
!
!Calculate Huber-type skipped mean
!
!****************************************
        subroutine huber_skip(hub,x)
            implicit none
        real(4) hub,x,r

        r=1.339
        if(abs(x)>=r) then
            hub = 0
        else
            hub = x
        end if
        return
        end
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
!****************************************
!
!Calculate Differentiate of Huber-type skipped mean
!
!****************************************
        subroutine diff_huber_skip(hub,x)
                implicit none
        real(4) hub,x,r
        r=1.339
        if(abs(x)>=r) then
                hub = 0
        else
                hub = 1
        end if
        return
        end

!****************************************
!
!Calculate Three part redescending estimator
!
!****************************************
        subroutine three_part(hub,x)
                implicit none
        real(4) hub,x,a,b,r
        real(4) tmpx
        integer sig

        a=1.7
        b=3.4
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
        r=8.5
        tmpx = abs(x)
        if(tmpx==0) then
                sig = 1
        else
                sig = x/tmpx
        end if
        if(tmpx>=r) then
                hub = 0
        elseif(tmpx>=b) then
                hub = ((a*sig)*(r-tmpx))/(r-b)
        elseif(tmpx>=a) then
                hub = a*sig
        else
                hub = x
        end if
        return
        end

!****************************************
!
!Calculate Differentiate of Three part redescending estimator
!
!****************************************
        subroutine diff_three_part(hub,x)
                implicit none
        real(4) hub,x,a,b,r
        real(4) tmpx
        integer sig
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
        a=1.7
        b=3.4
        r=8.5
        tmpx = abs(x)
        if(tmpx==0) then
                sig = 1
        else
                sig = x/tmpx
        end if
        if(tmpx>=r) then
                hub = 0
        elseif(tmpx>=b) then
                hub = (-a*sig)/(r-b)
        elseif(tmpx>=a) then
                hub = 0
        else
                hub = 1
        end if
        return
        end

!****************************************
!
!Calculate Tn : M-estimator
!
!****************************************
        subroutine m_est(me,med,x,xts,sn,n,sel)
                implicit none
        integer n,sel,i
        real(4) x(n),xts(n),sumhub,sumdiff
```

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

```
        real(4) sn,me,med,hub,diff
        sumhub=0
        sumdiff=0
        do i=1,size(x)
                if(sel==1) then !Huber estimator
                        call huber_est(hub,xts(i))
                        call diff_huber_est(diff,xts(i))
                elseif(sel==2) then !Huber-type skipped mean
                        call huber_skip(hub,xts(i))
                        call diff_huber_skip(diff,xts(i))
                else !Three part...
                        call three_part(hub,xts(i))
                        call diff_three_part(diff,xts(i))
                end if
                sumhub = sumhub+hub
                sumdiff = sumdiff+diff
        end do
        me = med+((sn*sumhub)/sumdiff)
        return
        end

!****************************************
!
!Calculate normal function
!
!****************************************
        subroutine norm_fun(nor,x)
                implicit none
        real nor,x
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
 nor = 0.3989423*EXP(-x*x/2)
        return
        end

!****************************************
!
!Calculate Tn of f-hat(x)
!
!****************************************
        subroutine Tn_h(tn,x,n,sel)
                implicit none
        integer n,i,j,sel
        real(4) x(n),tn,sd,sumfx
        real(4) h,nor,tmp,tmp1,sum1,sum2,fx(n)
        call std(sd,x,n)
        if(sel==1) then
                h=(2*sd)/(n**(0.2))
        elseif(sel==2) then
                h= sd/2
        else
                h= sd/4
        end if
        sum1=0
        sum2=0
        do i=1,size(x)
                sumfx=0
                do j=1,size(x)
                        tmp = (x(i)-x(j))/h
                        call norm_fun(nor,tmp)
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
                sumfx = sumfx+nor
        end do
        fx(i) = sumfx/(n*h)
        sum1 = sum1+fx(i)
end do
do i=1,size(x)
        tmp1 = (x(i)*fx(i))/sum1
        sum2 = sum2+tmp1
end do
tn = sum2
return
end
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ภาคผนวก ข

โปรแกรมสำหรับคำนวณค่าของ EIF

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

โปรแกรมสำหรับคำนวณค่าของ Empirical influence function (EIF) ของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนลที่เลือกค่า Window width เป็น $2s/n^{1/5}$ และเลือก Kernel function เป็น Gaussian kernel โดยกำหนดให้ข้อมูลตัวอย่างมีค่าผิดปกติ 1 ตัวรวมอยู่ด้วย และให้ค่าผิดปกติแปรค่าจาก –500 ไปถึง 500

1. ความหมายของตัวแปรในโปรแกรม

ตัวแปรในส่วนของโปรแกรมหลัก มีดังนี้

n	แทน ขนาดตัวอย่าง
tn(1000,2)	แทน ขนาดตัวอย่าง
rn(22)	แทน ตัวอย่างสุ่มโดยมีขนาดตัวอย่างตามที่กำหนด คือ 22 39 และ 100
den	แทน ค่าประมาณของตัวประมาณค่าเฉลี่ยจากค่าประมาณฟังก์ชันความหนาแน่น
mn	แทน ค่าเฉลี่ยตัวอย่าง หรือ ($\bar{X}$)
avr	แทน ค่าเฉลี่ยของประชากร
var	แทน ความแปรปรวนของประชากร

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ตัวแปรในส่วนของ subroutine rand2(avr,var,xout,xobs,sz) มีดังนี้

n2	แทน จำนวนข้อมูลที่เป็นปกติ
sz	แทน ขนาดตัวอย่าง ตามที่กำหนดซึ่งก็คือ 22 39 และ 100
xout	แทน ค่าผิดปกติ
x(sz)	แทน ตัวอย่างสุ่มที่สุ่มมาจากการแจกแจงแบบเอกรูป
xobs(sz)	แทน ตัวอย่างสุ่มที่สุ่มมาจากการแจกแจงแบบปกติ
avr	แทน ค่าเฉลี่ยของประชากร
var	แทน ความแปรปรวนของประชากร
zeed	แทน ค่าใดๆ ที่สุ่มมาจากการแจกแจงแบบเอกรูป มีค่าระหว่าง 0 ถึง 1
iseed	แทน ค่าเริ่มต้นสำหรับการสุ่มตัวอย่าง มีค่าระหว่าง 0 ถึง 2147483646

ตัวแปรในส่วนของ subroutine bbsort(x,n) มีดังนี้

n	แทน จำนวนข้อมูล
x(n)	แทน ค่าของข้อมูลจำนวน n ตัว
s	แทน ตัวแปรสำหรับการเปลี่ยนค่า

ตัวแปรในส่วนของ subroutine mean(mn,x,n) มีดังนี้

n	แทน ขนาดของตัวอย่างสุ่ม
x(n)	แทน ตัวอย่างสุ่มขนาด n
mn	แทน ค่าเฉลี่ยตัวอย่าง
sum	แทน ผลรวมของค่าจากตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine median(md,x,n) มีดังนี้

n	แทน ขนาดของตัวอย่างสุ่ม
pos	แทน ตำแหน่งกลางของข้อมูล
x(n)	แทน ตัวอย่างสุ่มขนาด n
md	แทน ค่ามัธยฐานของชุดตัวอย่างสุ่ม

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ตัวแปรในส่วนของ subroutine std(sd,x,n) มีดังนี้

n	แทน ขนาดของตัวอย่างสุ่ม
x(n)	แทน ตัวอย่างสุ่มขนาด n
sd	แทน ส่วนเบี่ยงเบนมาตรฐานของตัวอย่างสุ่ม
tmpmn	แทน ค่าเฉลี่ยของตัวอย่างสุ่ม
sum	แทน ผลรวมของค่าจากตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine norm_fun(nor,x) มีดังนี้

n	แทน ขนาดตัวอย่าง
nor	แทน ค่าของฟังก์ชันความหนาแน่นความน่าจะเป็น ณ จุด x
x	แทน ค่าของตัวอย่างสุ่ม

ตัวแปรในส่วนของ subroutine Tn_h(tn,x,n,sel) มีดังนี้

n	แทน ขนาดตัวอย่าง
nor	แทน ค่าของ kernel function ณ จุด x
x(n)	แทน ค่าของตัวอย่างสุ่มขนาด n
sel	แทน ตัวแปรที่ใช้เลือกค่าของ Window width
	โดย sel = 1 ถ้าต้องการเลือก Window width เป็น $2s/n^{1/5}$
	sel = 2 ถ้าต้องการเลือก Window width เป็น (1/2)s
	sel = 3 ถ้าต้องการเลือก Window width เป็น (1/4)s
tn	แทน ค่าประมาณของคัวประมาณค่าเฉลี่ยจากค่าประมาณฟังก์ชันความหนาแน่น
sd	แทน ส่วนเบี่ยงเบนมาตรฐานของค่าของตัวอย่างสุ่ม
sumfx	แทน ผลรวมของค่า Kernel function
h	แทน ค่าของ Window width
tmp	แทน ค่าที่แปลงจากค่าของตัวอย่างสุ่ม
tmp1	แทน ผลคูณของค่าสังเกต x กับความน่าจะเป็นที่ได้จากการประมาณฟังก์ชันความหนาแน่นความน่าจะเป็นแบบ Kernel ณ จุด x
sum1	แทน ผลรวมของค่า Kernel function
sum2	แทน ผลรวมของผลคูณของค่าสังเกต x กับความน่าจะเป็นที่ได้จากการประมาณฟังก์ชันความหนาแน่นความน่าจะเป็นแบบ Kernel ณ จุด x
fx(n)	แทน ค่าของฟังก์ชันประมาณความหนาแน่นความน่าจะเป็นแบบ Kernel ณ จุด x

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

2. ขั้นตอนการทำงาน

2.1 จำลองแบบข้อมูลจากการแจกแจงแบบปกติมาตรฐานขนาด 22 39 และ 100

2.2 สร้างค่าผิดปกติในชุดตัวอย่างให้แปรค่าตั้งแต่ –500 ถึง 500

2.3 ประมาณค่าเฉลี่ยจากการถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่น และค่าเฉลี่ยจากตัวอย่าง

3. แสดงรายละเอียดของโปรแกรม

```
!program empirical  influence  function
!implicit none
use  MSIMSL

integer n
real(4) tn(100,2)
real(4) rn(22),den,mn
real(4) avr,var
integer i

n=22
r=100
OPEN (3,FILE='.\output2.txt')
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
!write(3,*)'=====================Summary======================='

! input mean and variance for random sample
                print*,'Generate data from Contaminate Normal Dist.'
                print*,'with mean of gr.1'
                read*,avr
                print*,'variance of gr.1'
                read*,var
                do i=0,r
                call rand(avr,var,i,rn,size(rn))
                print *,rn
                !calculate den. by h(Tn5)
                call mean(mn,rn,size(rn))
                print *,'Tn (MEAN) = ',mn
```

```
                tn(i,1)=mn
                !calculate den. by h(Tn5)
                call Tn_h(den,rn,size(rn))
                print *,'Tn (density by h) = ',den
                tn(i,2)=den
                print *,'======================'
                write(3,1)tn(i,1),tn(i,2)
                end do

      1         format(F10.4,F10.4)
                end

!****************************************
!
!Generate random sample  from  contaminate normal  dist.  (with 1 outlier)
!
!****************************************
                subroutine rand(avr,var,xout,xobs,sz)
                        implicit none
                 integer i,n2,sz,xout
                 real(4) x(sz),xobs(sz)
                 real(4) avr,var,zeed,iseed

                call random_seed()
                call random_number(zeed)
                iseed = zeed * 2147483646
                CALL RNSET (iseed)
                CALL RNNOR (sz,x)
                n2 = sz-1
                        do 15 i=1,n2
```

มหาวิทยาลัยศิลปากร ลิขสิทธิ์

```
                    xobs(i) = (x(i)*(var**0.5)) + avr
15      continue
                    xobs(sz) = xout
                end
!****************************************
!
!Bubble sort (ascending order)
!
!****************************************
        subroutine bbsort(x,n)
                implicit none
        integer n,i,j
        real(4) x(n),s

        do 50 i=1,n-1
                do 50 j=1,n-i
                    if(x(j).GT.x(j+1)) then
                        s=x(j)
                        x(j)=x(j+1)
                        x(j+1)=s
                    end if
50      continue
        return
        end

!****************************************
!
!Calculate mean
!
!****************************************
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
        subroutine mean(mn,x,n)
                implicit none
        integer n,i
        real(4) x(n),mn,sum

        sum=0
        do i=1,size(x)
                sum = sum+x(i)
        end do
        mn = sum/size(x)
        return
        end

!****************************************
!
!Calculate standardivation
!
!****************************************
        subroutine std(sd,x,n)
                implicit none

        integer n,i
        real(4) x(n),sd,tmpmn
        real(4) sum

        call mean(tmpmn,x,n)
        do i=1,size(x)
                sum=sum+((x(i)-tmpmn)**2)
        end do
        sd = sqrt(sum/(size(x)-1))
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
        return
        end

!****************************************
!
!Calculate normal function
!
!****************************************
        subroutine norm_fun(nor,x)
                implicit none
        real nor,x

         nor = 0.3989423*EXP(-x*x/2)
        return
        end
!****************************************
!
!Calculate Tn of f-hat(x)
!
!****************************************
        subroutine Tn_h(tn,x,n)
                implicit none
        integer n,i,j
        real(4) x(n),tn,sd,sumfx
        real(4) h,nor,tmp,tmp1,sum1,sum2,fx(n)

        call std(sd,x,n)
        h=(2*sd)/(n**(0.2))
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

```
sum1=0
sum2=0
do i=1,size(x)
        sumfx=0
        do j=1,size(x)
                tmp = (x(i) - x(j))/h
                call norm_fun(nor,tmp)
                sumfx = sumfx+nor
        end do
        fx(i) = sumfx/(n*h)
        sum1 = sum1+fx(i)
end do
do i=1,size(x)
        tmp1 = (x(i)*fx(i))/sum1
        sum2 = sum2+tmp1
end do
tn = sum2
return
end
```

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์

ประวัติผู้วิจัย

ชื่อ - สกุล	นางสาวกนกกาญจน์ รัตนไพบูลย์
ที่อยู่	26 หมู่ 4 ต.โรงหีบ อ.บางคนที จ.สมุทรสงคราม
ประวัติการศึกษา	
พ.ศ. 2543	สำเร็จการศึกษาวิทยาศาสตรบัณฑิต สาขาสถิติ จากมหาวิทยาลัยศิลปากร พระราชวังสนามจันทร์ นครปฐม
พ.ศ. 2545	ศึกษาต่อระดับปริญญามหาบัณฑิต สาขาสถิติประยุกต์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

มหาวิทยาลัยศิลปากร สงวนลิขสิทธิ์