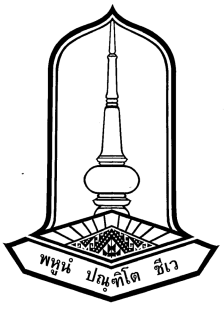


การพัฒนาวิธีการประมาณค่าข้อมูลสูญหาย

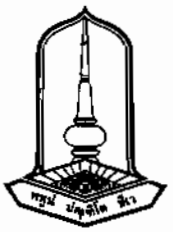
ไพทูรย์ มุลิวัลย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ
ตุลาคม 2556

การพัฒนาวิธีการประมาณค่าข้อมูลสูญหาย

ไพทูรย์ มุลิวัลย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ
ตุลาคม 2556



คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายไพทูรย์ มุลิวัลย์ แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....	ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ทรงศักดิ์ ภูสีอ่อน)	(อาจารย์บัณฑิตศึกษาภายนอกคณะ)
.....	กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นิภาพร ชุติมันต์)	(ประธานกรรมการควบคุมวิทยานิพนธ์)
.....	กรรมการ
(อาจารย์ ดร.ประภาส ผิวอ่อน)	(กรรมการควบคุมวิทยานิพนธ์)
.....	กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุนทรพจน์ ดำรงค์พานิช)	(อาจารย์บัณฑิตศึกษาภายนอกคณะ)
.....	กรรมการ
(อาจารย์ ดร.นิตยา บรรเทา)	(ผู้ทรงคุณวุฒิภายนอก)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ ของมหาวิทยาลัยมหาสารคาม

.....	.....
(ศาสตราจารย์ ดร.ละออศรี เสนาะเมือง)	(รศ.เทียนศักดิ์ เมฆพรรณโอภาส)
คณบดีคณะวิทยาศาสตร์	ผู้รักษาการคณบดีบัณฑิตวิทยาลัย

วันที่ 22 เดือน ต.ค. พ.ศ. 2556

วิทยานิพนธ์ฉบับนี้ ได้รับทุนอุดหนุนงานวิจัยสำหรับนิสิตระดับบัณฑิตศึกษา
มหาวิทยาลัยมหาสารคาม งบประมาณเงินรายได้ประจำปีงบประมาณ 2555

สารบัญ

หน้า

.

สารบัญตาราง

สารบัญภาพประกอบ

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการพัฒนาวิธีการประมาณค่าข้อมูลสูญหาย เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบ HDD-CIM กับวิธี HDD โดยใช้ข้อมูลทุติยภูมิ ซึ่งมีลำดับขั้นตอนการนำเสนอดังนี้

5.1 ความมุ่งหมายของการวิจัย

5.2 สรุปผล

5.3 อภิปรายผล

5.4 ข้อเสนอแนะ

5.1 ความมุ่งหมายของการวิจัย

การวิจัยครั้งนี้มีความมุ่งหมาย เพื่อสร้างวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบ HDD-CIM กับวิธี HDD (Hot-Deck) และวิธี CIM (Corrected –item Mean Substitution) จากขนาดตัวอย่างและร้อยละข้อมูลสูญหายต่างกันโดยพิจารณาเปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียง ความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ และศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย ในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน โดยใช้ข้อมูลทุติยภูมิ มีตัวแปรและขอบเขตที่ใช้ในการวิจัยดังต่อไปนี้

5.1.1 ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลทุติยภูมิ จากการสำรวจความคิดเห็นของประชาชนเกี่ยวกับสถานการณ์การแพร่ระบาดยาเสพติด พ.ศ. 2554 จาก 13 อำเภอ ที่สำรวจโดยสำนักงานสถิติจังหวัดมหาสารคาม จำนวน 2,975 ระเบียน เนื่องจากมีบางระเบียนที่ไม่สมบูรณ์ทำให้ได้จำนวนข้อมูลที่นำมาใช้ในการศึกษาครั้งนี้จำนวน 2,970 ระเบียน

5.1.2 การเปรียบเทียบประสิทธิภาพของวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM วิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีการประมาณข้อมูลสูญหายแบบ HDD

5.1.3 ตัวแปรที่ศึกษามีดังต่อไปนี้

5.1.3.1 ตัวแปรอิสระ

1) วิธีจัดการข้อมูลสูญหาย จำแนกเป็น 3 วิธี คือ
   (1) วิธีการประมาณข้อมูลสูญหายแบบ CIM
   (2) วิธีการประมาณข้อมูลสูญหายแบบ HDD
   (3) วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM
2) ขนาดตัวอย่าง
   (1) ขนาดตัวอย่าง 100
   (2) ขนาดตัวอย่าง 200

(3) ขนาดตัวอย่าง 500

3) ร้อยละข้อมูลสูญหาย

(1) ร้อยละ 5

(2) ร้อยละ 10

(3) ร้อยละ 15

(4) ร้อยละ 20

(5) ร้อยละ 30

5.1.3.2 ตัวแปรตาม

1) ประสิทธิภาพของวิธีการประมาณข้อมูลสูญหาย

2) ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง และร้อยละข้อมูลสูญหาย

5.2 สรุปผล

5.3.1 ผลการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM พบว่า วิธีการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียง ความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ต่ำ นั่นแสดงว่า วิธีการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นมีประสิทธิภาพในการประมาณค่าข้อมูลสูญหายดีกว่าวิธีที่นำมาเปรียบเทียบ

5.3.2 ผลการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย สรุปได้ดังต่อไปนี้

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย ที่ระดับนัยสำคัญ .05 สรุปผลได้ดังต่อไปนี้ เมื่อขนาดตัวอย่าง 100 200 และ 500 การประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพสูงที่สุด

5.3.3 ผลการศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย ที่มีผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน สรุปได้ดังต่อไปนี้

5.3.3.1 ผลการศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ที่ระดับนัยสำคัญ .05 สรุปได้ว่า เมื่อใช้ขนาดตัวอย่างต่างกัน วิธีการจัดการข้อมูลสูญหายต่างกันส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

5.3.3.2 ผลการศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความเอนเอียง ที่ระดับนัยสำคัญ .05 สรุปได้ว่า เมื่อใช้ขนาดตัวอย่างต่างกัน วิธีการจัดการข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความเอนเอียง

5.3.3.3 ผลการศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความแปรปรวน ที่ระดับนัยสำคัญ .05 สรุปได้ว่า เมื่อใช้ขนาดตัวอย่างต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน เมื่อใช้ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

1) การใช้ขนาดตัวอย่างต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

2) การใช้ร้อยละข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

3) วิธีการจัดการข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

4) การใช้ขนาดตัวอย่างต่างกัน ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

5.3 อภิปรายผล

การวิจัยครั้งนี้ได้ข้อค้นพบเกี่ยวกับ วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ที่ผู้วิจัยพัฒนาขึ้นเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเลื่อนกำลังสองเฉลี่ย ความเอนเอียง และความแปรปรวน จากขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่แตกต่างกัน โดยใช้ข้อมูลทุติยภูมิกับวิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีการประมาณข้อมูลสูญหายแบบ HDD และศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย ในการประมาณค่าความคลาดเคลื่อนกำลังสอง ความเอนเอียงและความแปรปรวน ข้อค้นพบอภิปรายผลได้ดังต่อไปนี้

วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM มีความแกร่งต่อขนาดตัวอย่างและร้อยละข้อมูลสูญหาย และมีประสิทธิภาพสูงเมื่อใช้กับข้อมูลจริงที่มีระดับการวัดในมาตรานามบัญญัติ (Nominal Scale) และมาตราเรียงอันดับ (Ordinal Scale) ในทุก ๆ ขนาดตัวอย่างและร้อยละข้อมูลสูญหายทุกระดับ เนื่องจากได้นำวิธีการประมาณข้อมูลสูญหายแบบ HDD เป็นวิธีที่ให้ค่าความเอนเอียงต่ำสุดและมีความแม่นยำสูงสุดซึ่งสอดคล้องกับงานวิจัยของ Kevin Strike และคณะ (2001: 890-908) ที่ได้ทดลองโดยใช้วิธีแทนที่ข้อมูลสูญหายรวมทั้งสิ้น 10 วิธี โดยใช้ข้อมูลจาก National Health and Nutrition Examination Survey (NHANES) จำนวน 800 ตัวอย่าง และวิธีที่ดีที่สุดจะต้องมีค่าความเอนเอียงต่ำและมีความแม่นยำสูง พบว่า วิธี Hot Deck ที่ใช้การหาระยะทางแบบยูคลิด (Euclidean Distance) และคะแนนมาตรฐาน (Z-score Standardization) เป็นวิธีที่ให้ค่าความเอนเอียงต่ำสุดและมีความแม่นยำสูงสุด และงานวิจัยของ Andridge and Little (2010: 40 - 64) ที่ได้ทดลองโดยใช้ข้อมูลจาก The Third National Health and Nutrition Examination Survey (NHANES III) พบว่า วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพสูงเมื่อใช้กับข้อมูลจริงที่ตัวอย่างมีขนาดใหญ่และร้อยละข้อมูลสูญหายมีจำนวนมาก และ

ใช้วิธีการประมาณข้อมูลสูญหายแบบ CIM ที่มีคุณสมบัติเป็นตัวประมาณข้อมูลสูญหายที่ดีซึ่งสอดคล้องกับงานวิจัยของ Huisman (2000: 331 - 351) ได้ศึกษาวิธีแทนที่ข้อมูลสูญหายโดยใช้ข้อมูลจาก 4 ชุดข้อมูล ตัวอย่างมี 3 ขนาด โดยนำวิธีประมาณค่าข้อมูลสูญหายทั้ง 9 วิธีมาเปรียบเทียบกัน พบว่า วิธี CIM เป็นวิธีที่ดีที่สุดเนื่องจากให้ค่า Root Mean Square Deviation (RMSD) ต่ำที่สุด มาร่วมในการประมาณข้อมูลสูญหาย จึงทำให้วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM มีความแกร่งต่อขนาดตัวอย่างและร้อยละข้อมูลสูญหาย และมีประสิทธิภาพสูงในการประมาณค่าข้อมูลสูญหาย

5.4 ข้อเสนอแนะ

จากการพัฒนาวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM และเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียง และความแปรปรวนของวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM วิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีการประมาณข้อมูลสูญหายแบบ HDD จากขนาดตัวอย่างและร้อยละข้อมูลสูญหาย โดยใช้ข้อมูลทุติยภูมิ และศึกษาปฏิสัมพันธ์ระหว่างวิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน โดยใช้ข้อมูลทุติยภูมิ ผู้วิจัยมีข้อเสนอแนะดังต่อไปนี้

5.4.1 ข้อเสนอแนะในการนำผลงานการวิจัยไปใช้

วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM จะมีประสิทธิภาพสูงสุดในบางสถานการณ์ ดังนั้นการเลือกใช้จะขึ้นอยู่กับเงื่อนไขต่าง ๆ ที่เหมาะสม จากผลการวิจัยมีข้อเสนอแนะสำหรับการนำวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ไปใช้ ดังต่อไปนี้

วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM จะมีประสิทธิภาพสูงสุดเมื่อข้อมูลมีระดับการวัดเป็นมาตรานามบัญญัติ (Nominal Scale) และมาตราเรียงอันดับ (Ordinal Scale) ประเภทข้อมูลสูญหายแบบสุ่มสมบูรณ์ ใช้การชักตัวอย่างแบบสุ่มเชิงเดียวและเมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 มีร้อยละข้อมูลสูญหายที่ระดับ 5 10 15 20 และ 30

5.4.2 ข้อเสนอแนะในการทำวิจัยครั้งต่อไป

การพัฒนาวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ได้ทำการศึกษาภายใต้ข้อมูลสูญหายแบบสุ่มสมบูรณ์ การสุ่มตัวอย่างอย่างง่าย ดังนั้นผู้วิจัยมีข้อเสนอแนะในการทำการวิจัยครั้งต่อไป ดังนี้

5.4.2.1 ศึกษาภายใต้ประเภทการสูญหายแบบสุ่ม (MAR) และแบบไม่สุ่ม (NMAR)

5.4.2.2 ศึกษาภายใต้การสุ่มตัวอย่างแบบมีระบบ (Systematic Random Sampling) การสุ่มตัวอย่างแบบชั้นภูมิ (Stratified Random Sampling) การสุ่มตัวอย่างแบบกลุ่ม (Cluster Random Sampling)

5.4.2.3 ศึกษาวิธีการประมาณข้อมูลสูญหายที่นักวิจัยพัฒนาขึ้น กับวิธีการประมาณข้อมูลสูญหายแบบอื่น ๆ เช่น การตัดข้อมูลสูญหายแบบลิสท์ไวส์ วิธีการประมาณข้อมูลสูญหายด้วยการถดถอย วิธีการประมาณข้อมูลสูญหายแบบอีพีอาร์

บทที่ 4

ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้มีความมุ่งหมายเพื่อพัฒนาวิธีการประมาณข้อมูลสูญหายด้วยวิธี HDD-CIM และเปรียบเทียบประสิทธิภาพกับวิธีการประมาณข้อมูลสูญหายด้วยวิธี CIM และวิธีการประมาณข้อมูลสูญหายด้วยวิธี HDD จากขนาดตัวอย่างและร้อยละของข้อมูลสูญหายที่แตกต่างกัน โดยพิจารณาเปรียบเทียบการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน โดยใช้ข้อมูลจริง และศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหายขนาดตัวอย่างและร้อยละข้อมูลสูญหาย

4.1 สัญลักษณ์ที่ใช้ในการนำเสนอผลการวิเคราะห์ข้อมูล

CIM	หมายถึง	วิธีการประมาณข้อมูลสูญหายด้วยวิธี CIM
HDD	หมายถึง	วิธีการประมาณข้อมูลสูญหายด้วยวิธี Hot Deck
HDD-CIM	หมายถึง	วิธีการประมาณข้อมูลสูญหายด้วยวิธี HDD-CIM
p	หมายถึง	ค่าความน่าจะเป็นน้อยที่สุดที่จะปฏิเสธสมมติฐานหลัก เมื่อสมมติฐานหลักเป็นจริง

4.2 ลำดับขั้นตอนในการนำเสนอผลการวิเคราะห์ข้อมูล

ผลการวิเคราะห์ข้อมูลในการวิจัยครั้งนี้ จะนำเสนอเป็นตอน ๆ ดังนี้

ตอนที่ 1 ผลการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM

ตอนที่ 2 ผลการเปรียบเทียบประสิทธิภาพในการประมาณค่าข้อมูลสูญหายแบบใหม่เทียบกับวิธีการ HDD และวิธีการ CIM

4.3 ผลการวิเคราะห์ข้อมูล

ตอนที่ 1 ผลการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM

จากการศึกษาเอกสาร แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้องกับวิธีการจัดการข้อมูลสูญหายและวิเคราะห์แนวคิดของวิธีการประมาณค่าข้อมูลสูญหายแบบต่าง ๆ โดยพิจารณาจุดเด่นจุดด้อยของแต่ละวิธี เพื่อเป็นแนวทางในการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบ HDD-CIM พบว่า

1. ค่าประมาณค่าข้อมูลสูญหายในขั้นตอนที่หนึ่ง ที่นำเอาทั้งผลกระทบจากบุคคลและผลกระทบจากคำถามมาร่วมเป็นตัวถ่วงน้ำหนักเพื่อใช้สร้างชุดข้อมูลที่สมบูรณ์ชั่วคราว ทำให้ได้ตัวประมาณที่มีค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ต่ำ

2. ขั้นตอนที่สอง ที่หาค่าระยะทางระหว่างหน่วยสำรวจนั้น เพื่อหาระยะทางระหว่างคู่สำรวจ โดยที่คู่ใดมีระยะห่างกันน้อยแสดงว่าหน่วยสำรวจคู่นั้นมีคุณสมบัติใกล้กันมาก

3. ในขั้นตอนที่สาม ในการเลือกผู้ให้ข้อมูลที่มีหน่วยสำรวจที่สมบูรณ์แทนที่ข้อมูลสูญหาย ทำให้การประมาณค่าข้อมูลสูญหายมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียง ความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ต่ำ

ตอนที่ 2 ผลการเปรียบเทียบประสิทธิภาพในการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นเทียบกับวิธีการ HDD และวิธีการ CIM

การนำเสนอผลการเปรียบเทียบประสิทธิภาพในการประมาณค่าข้อมูลสูญหาที่พัฒนาขึ้นเทียบกับวิธีการ HDD และวิธีการ CIM ผู้วิจัยได้นำเสนอตามลำดับดังนี้

1. เปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

2. เปรียบเทียบค่าความเอนเอียง จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

3. เปรียบเทียบค่าความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

4. เปรียบเทียบค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

5. ผลการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ค่าความเอนเอียง ค่าความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย

รายละเอียดดังแสดงในตาราง 4.1 – 4.5

การนำเสนอผลการศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย ที่มีผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน ผู้วิจัยได้นำเสนอตามลำดับดังนี้

1. ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

2. ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความเอนเอียง

3. ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความแปรปรวน

รายละเอียดดังแสดงในตาราง 4.6 – 4.8

1. เปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ผลการเปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย แสดงดังตาราง 4.1

ตาราง 4.1 เปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ขนาดตัวอย่าง	ร้อยละข้อมูลสูญหาย	วิธีการจัดการข้อมูลสูญหาย		
		CIM	HDD	HDD-CIM
100	5	2.84612*	2.98794*	1.86727
	10	2.73195*	2.99894*	1.81650
	15	2.78234*	3.12866*	2.05583
	20	2.86297*	2.87705*	1.87539
	30	2.95501*	2.86625*	1.91216
200	5	3.07550	4.30629*	2.17792
	10	3.15447*	3.47519*	2.14103
	15	3.04765*	3.38747*	2.11127
	20	3.17141*	3.46227*	2.18833
	30	3.15519*	3.39253*	2.08301
500	5	2.97879*	3.31609*	1.95161
	10	3.07663*	3.40567*	2.06833
	15	3.01320*	3.23201*	2.06946
	20	2.88565*	3.08062*	1.82702
	30	3.07524*	3.12400*	1.86772

ตัวเลขขีดเส้นใต้ หมายถึง มีประสิทธิภาพสูงสุดในแต่ละระดับร้อยละข้อมูลสูญหาย

* หมายถึง ค่าประสิทธิภาพแตกต่างจากค่าประสิทธิภาพที่สูงสุด ที่ระดับนัยสำคัญ .05

จากตาราง 4.1 พบว่า เมื่อใช้ขนาดตัวอย่าง 100 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด

เมื่อใช้ตัวอย่างขนาด 200 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด

เมื่อใช้ขนาดตัวอย่าง 500 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกตามขนาดตัวอย่างและร้อยละข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพด้วยการวิเคราะห์ความแปรปรวน และทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพรายคู่ โดยวิธีเชฟเฟ่ สรุปได้ดังต่อไปนี้

เมื่อขนาดตัวอย่าง 100 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 200 ข้อมูลสูญหายร้อยละ 5 วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ที่ระดับนัยสำคัญ .05 และข้อมูลสูญหายร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 500 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ที่ระดับนัยสำคัญ .05

2. เปรียบเทียบค่าความเอนเอียง จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ผลการเปรียบเทียบค่าความเอนเอียง จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย แสดงดังตาราง 4.2

ตาราง 4.2 เปรียบเทียบค่าความเอนเอียง จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ขนาดตัวอย่าง	ร้อยละข้อมูลสูญหาย	วิธีการจัดการข้อมูลสูญหาย		
		CIM	HDD	HDD-CIM
100	5	1.64896*	1.69916*	1.33178
	10	1.63695*	1.71379*	1.33426
	15	1.65706*	1.75709*	1.41722
	20	1.68300*	1.68670*	1.35941
	30	1.71357*	1.68690*	1.37678

ตาราง 4.2 (ต่อ)

ขนาดตัวอย่าง	ร้อยละข้อมูลสูญหาย	วิธีการจัดการข้อมูลสูญหาย		
		CIM	HDD	HDD-CIM
200	5	1.73341*	1.90945*	1.45593
	10	1.76766*	1.85914*	1.45538
	15	1.73979*	1.83519*	1.44634
	20	1.77790*	1.85778*	1.47714
	30	1.77419*	1.83873*	1.44139
500	5	1.72006*	1.81582*	1.39149
	10	1.75027*	1.84328*	1.43256
	15	1.73365*	1.79480*	1.43133
	20	1.69728*	1.74292*	1.34833
	30	1.75240*	1.73756*	1.35369

ตัวเลขขีดเส้นใต้ หมายถึง มีประสิทธิภาพสูงสุดในแต่ละระดับร้อยละข้อมูลสูญหาย

* หมายถึง ค่าประสิทธิภาพแตกต่างจากค่าประสิทธิภาพที่สูงสุด ที่ระดับนัยสำคัญ .05

จากตาราง 4.2 พบว่า เมื่อใช้ขนาดตัวอย่าง 100 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความเอนเอียงต่ำที่สุด

เมื่อใช้ตัวอย่างขนาด 200 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความเอนเอียงที่สุด

เมื่อใช้ขนาดตัวอย่าง 500 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความเอนเอียงต่ำที่สุด

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความเอนเอียง จำแนกตามขนาดตัวอย่างและร้อยละข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพด้วยการวิเคราะห์ความแปรปรวน และทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพรายคู่ โดยวิธีเชฟเฟ่ สรุปได้ดังต่อไปนี้

เมื่อขนาดตัวอย่าง 100 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความเอนเอียง ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 200 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความเอนเอียง ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 500 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความเอนเอียง ที่ระดับนัยสำคัญ .05

3. เปรียบเทียบค่าความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ผลการเปรียบเทียบค่าความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย แสดงดังตาราง 4.3

ตาราง 4.3 เปรียบเทียบค่าความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ขนาดตัวอย่าง	ร้อยละข้อมูลสูญหาย	วิธีการจัดการข้อมูลสูญหาย		
		CIM	HDD	HDD-CIM
100	5	1.43571*	1.43904*	1.42064
	10	1.44989*	1.45872*	1.42018
	15	1.46457*	1.47769*	1.42377
	20	1.47011*	1.47575*	1.41166
	30	1.49382*	1.49852*	1.39973
200	5	1.38107	1.39880*	1.36766
	10	1.39258*	1.40699*	1.36325
	15	1.40628*	1.43078*	1.36512
	20	1.42071*	1.44842*	1.36430
	30	1.43212*	1.46643*	1.35092
500	5	1.39870*	1.40503*	1.38037
	10	1.40870*	1.42383*	1.37588
	15	1.42203*	1.44107*	1.37753
	20	1.43468*	1.45558*	1.37347
	30	1.44468*	1.48498*	1.36726

ตัวเลขขีดเส้นใต้ หมายถึง มีประสิทธิภาพสูงสุดในแต่ละระดับร้อยละข้อมูลสูญหาย

* หมายถึง ค่าประสิทธิภาพแตกต่างจากค่าประสิทธิภาพที่สูงสุด ที่ระดับนัยสำคัญ .05

จากตาราง 4.3 พบว่า เมื่อใช้ขนาดตัวอย่าง 100 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความแปรปรวนต่ำที่สุด

เมื่อใช้ตัวอย่างขนาด 200 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความแปรปรวนที่สุด

เมื่อใช้ขนาดตัวอย่าง 500 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าความแปรปรวนต่ำที่สุด

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความแปรปรวน จำแนกตามขนาดตัวอย่างและร้อยละข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพด้วยการวิเคราะห์ความแปรปรวน และทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพรายคู่ โดยวิธีเชฟเฟ่ สรุปได้ดังต่อไปนี้

เมื่อขนาดตัวอย่าง 100 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความแปรปรวน ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 200 ข้อมูลสูญหายร้อยละ 5 วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความแปรปรวน ที่ระดับนัยสำคัญ .05 และข้อมูลสูญหายร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความแปรปรวน ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 500 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุดในการประมาณค่าความแปรปรวน ที่ระดับนัยสำคัญ .05

4. เปรียบเทียบค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่างร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ผลการเปรียบเทียบค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย แสดงดังตาราง 4.4

ตาราง 4.4 เปรียบเทียบค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย

ขนาดตัวอย่าง	ร้อยละข้อมูลสูญหาย	วิธีการจัดการข้อมูลสูญหาย		
		CIM	HDD	HDD-CIM
100	5	42.53*	39.52*	32.66
	10	43.46*	40.74*	35.98
	15	43.42*	40.91*	36.30
	20	42.03*	39.19*	33.55
	30	43.94*	39.44*	35.37
200	5	44.04*	39.93*	33.40
	10	43.08*	39.34*	33.29
	15	42.25*	39.68*	33.45
	20	43.21*	39.20*	33.57
	30	43.62*	39.96*	33.88
500	5	43.33*	41.70*	32.97
	10	43.28*	41.07*	33.72
	15	43.87*	39.41*	34.22
	20	42.57*	36.88*	30.28
	30	43.34*	36.29*	30.59

ตัวเลขขีดเส้นใต้ หมายถึง มีประสิทธิภาพสูงสุดในแต่ละระดับร้อยละข้อมูลสูญหาย

* หมายถึง ค่าประสิทธิภาพแตกต่างจากค่าประสิทธิภาพที่สูงสุด ที่ระดับนัยสำคัญ .05

จากตาราง 4.4 พบว่า เมื่อใช้ขนาดตัวอย่าง 100 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ต่ำที่สุด

เมื่อใช้ตัวอย่างขนาด 200 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ที่สุด

เมื่อใช้ขนาดตัวอย่าง 500 ที่ข้อมูลสูญหายร้อยละ 5 ถึงร้อยละ 30 พบว่า วิธีประมาณค่าข้อมูลสูญหายด้วยวิธีการ HDD-CIM มีประสิทธิภาพมากที่สุดเนื่องจากมีค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ต่ำที่สุด

จากการเปรียบเทียบค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่างและร้อยละข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพด้วยการวิเคราะห์ความแปรปรวน และทดสอบความแตกต่างของค่าเฉลี่ยประสิทธิภาพรายคู่ โดยวิธีเชฟเฟ่ สรุปได้ดังต่อไปนี้

เมื่อขนาดตัวอย่าง 100 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุด ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 200 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุด ที่ระดับนัยสำคัญ .05

เมื่อขนาดตัวอย่าง 500 ข้อมูลสูญหายร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 20 และร้อยละ 30 วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM ซึ่งเป็นวิธีการที่มีประสิทธิภาพสูงที่สุด ที่ระดับนัยสำคัญ .05

5. ผลการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ค่าความเอนเอียง ค่าความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย แสดงดังตาราง 4.5

ตาราง 4.5 เปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ค่าความเอนเอียง ค่าความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย

การประมาณค่า	ขนาดตัวอย่าง	วิธีการจัดการข้อมูลสูญหาย		
		CIM	HDD	HDD-CIM
ความคลาดเคลื่อนกำลังสองเฉลี่ย	100	2.83568*	2.97177*	<u>1.90543</u>
	200	3.09597*	3.68108*	<u>2.14829</u>
	500	2.99506*	3.26545*	<u>1.95474</u>
ความเอนเอียง	100	1.66791*	1.70873*	<u>1.36389</u>
	200	1.74747*	1.87281*	<u>1.45383</u>
	500	1.72646*	1.79845*	<u>1.39148</u>
ความแปรปรวน	100	1.46282*	1.46994*	<u>1.41520</u>
	200	1.39544*	1.41583*	<u>1.36484</u>
	500	1.41253*	1.42727*	<u>1.37709</u>
ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์	100	43.08*	39.96*	<u>34.77</u>
	200	43.29*	39.64*	<u>33.51</u>
	500	43.34*	39.65*	<u>32.49</u>

ตัวเลขขีดเส้นใต้ หมายถึง มีประสิทธิภาพสูงสุดในแต่ละขนาดตัวอย่าง

* หมายถึง ค่าประสิทธิภาพแตกต่างจากค่าประสิทธิภาพที่สูงสุด ที่ระดับนัยสำคัญ .05

จากตาราง 4.5 พบว่า ในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพสูงที่สุด

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยด้วยการวิเคราะห์ความแปรปรวน พบว่า เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีประมาณข้อมูลสูญหายแบบ HDD ที่ระดับนัยสำคัญ .05 ($p < .000$)

ในการประมาณค่าความเอนเอียง เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพสูงที่สุด

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความเอนเอียง จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าความเอนเอียงด้วยการวิเคราะห์ความแปรปรวน พบว่า เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีประมาณข้อมูลสูญหายแบบ HDD ที่ระดับนัยสำคัญ .05 ($p < .000$)

ในการประมาณค่าความแปรปรวน เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพสูงที่สุด

จากการเปรียบเทียบประสิทธิภาพการประมาณค่าความแปรปรวน จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าความแปรปรวนด้วยการวิเคราะห์ความแปรปรวน พบว่า เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีประมาณข้อมูลสูญหายแบบ HDD ที่ระดับนัยสำคัญ .05 ($p < .000$)

จากค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพสูงที่สุด

จากการเปรียบเทียบประสิทธิภาพจากค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย โดยการทดสอบความแตกต่างของค่าความแปรปรวนด้วยการวิเคราะห์ความแปรปรวน พบว่า เมื่อใช้ขนาดตัวอย่าง 100 200 และ 500 วิธีประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพแตกต่างแตกต่างจากวิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีประมาณข้อมูลสูญหายแบบ HDD ที่ระดับนัยสำคัญ .05

การนำเสนอผลการศึกษาปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย ที่มีผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยความเอนเอียงและความแปรปรวน

1. ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่าง วิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย แสดงดังตาราง 4.6

ตาราง 4.6 ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่าง วิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

Source of Variation	Sum of Squares	df	Mean Square	*F*	p - value
อิทธิพลหลัก					
ขนาดตัวอย่าง	45.291	2	22.646	13.995*	.000
ร้อยละข้อมูลสูญหาย	2.647	4	.622	.409	.802
วิธีการจัดการข้อมูลสูญหาย	381.836	2	190.918	117.988*	.000
ปฏิสัมพันธ์สองปัจจัย					
ขนาดตัวอย่าง x ร้อยละข้อมูลสูญหาย	9.562	8	1.1915	.739	.657
ขนาดตัวอย่าง x วิธีการจัดการข้อมูลสูญหาย	8.293	4	2.073	1.281	.275
ร้อยละข้อมูลสูญหาย x วิธีการจัดการข้อมูลสูญหาย	7.724	8	.966	.597	.781
ปฏิสัมพันธ์สามปัจจัย					
ขนาดตัวอย่าง x ร้อยละข้อมูลสูญหาย x วิธีการจัดการข้อมูลสูญหาย	10.695	16	.668	.413	.980
ความคลาดเคลื่อน	4126.177	2550	1.618		
รวม	24005.646	2595			

* มีนัยสำคัญทางสถิติที่ระดับ .05

จากตาราง 4.6 พบว่า ไม่มีปฏิสัมพันธ์สามปัจจัยระหว่างขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย นั่นคือการใช้ขนาดตัวอย่างต่างกัน ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

เมื่อพิจารณาปฏิสัมพันธ์สองปัจจัย พบว่า ไม่มีปฏิสัมพันธ์สองปัจจัยระหว่าง 1) ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย 2) ขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย และ 3) ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย นั่นคือ 1) เมื่อใช้ขนาดตัวอย่างต่างกันและร้อยละข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย 2) เมื่อใช้ขนาดตัวอย่างต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย 3) เมื่อใช้ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

เมื่อพิจารณาอิทธิพลหลัก พบว่า ขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหายมีนัยสำคัญที่ระดับสำคัญ .05 นั่นคือ การใช้ขนาดตัวอย่างต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ดังตาราง 4.7-4.8

ตาราง 4.7 ผลการเปรียบเทียบพหุคูณ (Multiple Comparison) ด้วยวิธีการของเชฟเฟ่ จำแนกตามขนาดตัวอย่าง ที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

ขนาดตัวอย่าง		ผลต่างของค่าเฉลี่ย	p – value
100	200	-0.3286*	.000
	500	-0.1675*	.003
200	500	0.1611*	.012

* มีนัยสำคัญทางสถิติที่ระดับ .05

จากตาราง 4.7 พบว่า ตัวอย่างขนาด 100 200 และ 500 จะส่งผลต่อประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่างกัน ที่ระดับนัยสำคัญ .05 โดยที่ตัวอย่างขนาด 100 จะมีประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด

ตาราง 4.8 ผลการเปรียบเทียบพหุคูณ (Multiple Comparison) ด้วยวิธีการของเชฟเฟ่ จำแนกตามวิธีการประมาณค่าข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

วิธีการ		ผลต่างของค่าเฉลี่ย	p – value
CIM	HDD	-0.2172*	.000
	HDD-CIM	0.9510*	.000
HDD	HDD-CIM	1.1682*	.000

* มีนัยสำคัญทางสถิติที่ระดับ .05

จากตาราง 4.8 พบว่า การใช้วิธีการประมาณค่าข้อมูลสูญหายต่างกันจะส่งผลต่อประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่างกัน ที่ระดับนัยสำคัญ .05 โดยที่วิธีการประมาณค่าข้อมูลสูญหายแบบ HDD-CIM จะมีประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด

2. ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความเอนเอียง

ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่าง วิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่มีต่อประสิทธิภาพการประมาณค่าความเอนเอียง แสดงดังตาราง 4.9

ตาราง 4.9 ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่าง วิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่มีต่อประสิทธิภาพการประมาณค่าความเอนเอียง

Source of Variation	Sum of Squares	df	Mean Square	*F*	p - value
อิทธิพลหลัก					
ขนาดตัวอย่าง	4.270	2	2.135	38.393*	.000
ร้อยละข้อมูลสูญหาย	.073	4	.018	.327	.860
วิธีการจัดการข้อมูลสูญหาย	36.571	2	18.286	328.795*	.000
ปฏิสัมพันธ์สองปัจจัย					
ขนาดตัวอย่าง x ร้อยละข้อมูลสูญหาย	.670	8	.084	1.506	.150
ปฏิสัมพันธ์สองปัจจัย					
ขนาดตัวอย่าง x วิธีการจัดการข้อมูลสูญหาย	.334	4	.084	1.503	.199
ร้อยละข้อมูลสูญหาย x วิธีการจัดการข้อมูลสูญหาย	.384	8	.048	.862	.548
ปฏิสัมพันธ์สามปัจจัย					
ขนาดตัวอย่าง x ร้อยละข้อมูลสูญหาย x วิธีการจัดการข้อมูลสูญหาย	.270	16	.017	.303	.996
ความคลาดเคลื่อน	141.816	2550	.056		

* มีนัยสำคัญทางสถิติที่ระดับ .05

จากตาราง 4.9 พบว่า ไม่มีปฏิสัมพันธ์สามปัจจัยระหว่างขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย นั่นคือการใช้ขนาดตัวอย่างต่างกัน ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความเอนเอียง

เมื่อพิจารณาปฏิสัมพันธ์สองปัจจัย พบว่า ไม่มีปฏิสัมพันธ์สองปัจจัยระหว่าง 1) ขนาดตัวอย่างและร้อยละข้อมูลสูญหาย 2) ขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย และ 3) ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย นั่นคือ 1) เมื่อใช้ขนาดตัวอย่างต่างกันและร้อยละข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย 2) เมื่อใช้ขนาดตัวอย่างต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย 3) เมื่อใช้ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความเอนเอียง

เมื่อพิจารณาอิทธิพลหลัก พบว่า ขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหายมีนัยสำคัญที่ระดับสำคัญ .05 นั่นคือ การใช้ขนาดตัวอย่างต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ส่งผลต่อประสิทธิภาพการประมาณค่าความเอนเอียง แสดงดังตาราง 4.10 – 4.11

ตาราง 4.10 ผลการเปรียบเทียบพหุคูณ (Multiple Comparison) ด้วยวิธีการของเชฟเฟ่ จำแนกตามขนาดตัวอย่าง ที่มีต่อประสิทธิภาพการประมาณค่าความเอนเอียง

ขนาดตัวอย่าง		ผลต่างของค่าเฉลี่ย	p – value
100	200	-0.0995*	.000
	500	-0.0588*	.003
200	500	0.0407	.065

* มีนัยสำคัญทางสถิติที่ระดับ .05

จากตาราง 4.10 พบว่า การใช้ขนาดตัวอย่างต่างกันจะส่งผลต่อประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่างกัน ที่ระดับนัยสำคัญ .05 โดยที่ตัวอย่างขนาด 100 จะมีประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่างจากตัวอย่างขนาด 200 และ 500 ที่ระดับนัยสำคัญ .05 และตัวอย่างขนาด 100 จะมีประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด

ตาราง 7.11 ผลการเปรียบเทียบพหุคูณ (Multiple Comparison) ด้วยวิธีการของเชฟเฟ่ จำแนกตามวิธีการประมาณค่าข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

วิธีการ		ผลต่างของค่าเฉลี่ย	p – value
CIM	HDD	-0.0468*	.027
	HDD-CIM	0.3279*	.000
HDD	HDD-CIM	0.3748*	.000

* มีนัยสำคัญทางสถิติที่ระดับ .05

จากตาราง 4.11 พบว่า การใช้วิธีการประมาณค่าข้อมูลสูญหายต่างกันจะส่งผลต่อประสิทธิภาพในการประมาณค่าความเอนเอียงต่างกัน ที่ระดับนัยสำคัญ .05 โดยที่วิธีการประมาณค่าข้อมูลสูญหายแบบ HDD-CIM จะมีประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนดีที่สุด

3. ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าความแปรปรวน

ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่าง วิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่มีต่อประสิทธิภาพการประมาณค่าความแปรปรวน แสดงดังตาราง 4.12

ตาราง 4.12 ผลการวิเคราะห์ปฏิสัมพันธ์ระหว่าง วิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่มีต่อประสิทธิภาพการประมาณค่าความแปรปรวน

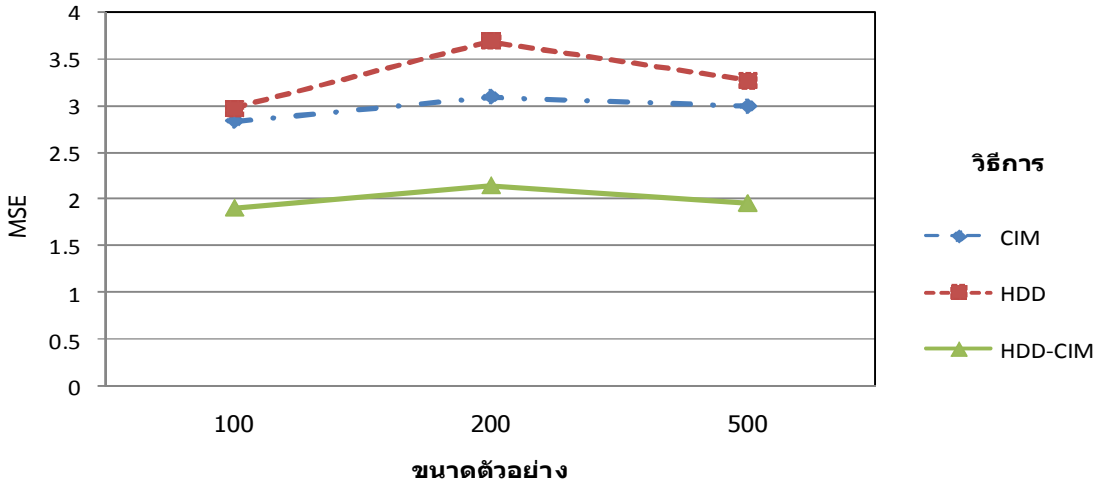
Source of Variation	Sum of Squares	df	Mean Square	*F*	p - value
อิทธิพลหลัก					
ขนาดตัวอย่าง	.956	2	.478	600.262*	.000
ร้อยละข้อมูลสูญหาย	.262	4	.066	82.377*	.000
วิธีการจัดการข้อมูลสูญหาย	.943	2	.472	592.311*	.000
ปฏิสัมพันธ์สองปัจจัย					
ขนาดตัวอย่าง x ร้อยละข้อมูลสูญหาย	.010	8	.001	1.624	.113
ขนาดตัวอย่าง x วิธีการจัดการข้อมูลสูญหาย	.022	4	.005	6.765*	.000
ร้อยละข้อมูลสูญหาย x วิธีการจัดการข้อมูลสูญหาย	.254	8	.032	39.910*	.000
ปฏิสัมพันธ์สามปัจจัย					
ขนาดตัวอย่าง x ร้อยละข้อมูลสูญหาย x วิธีการจัดการข้อมูลสูญหาย	.009	16	.001	.123	.733
ความคลาดเคลื่อน	2.031	2550	.001		
รวม	5291.035	2595			

* มีนัยสำคัญทางสถิติที่ระดับ .05

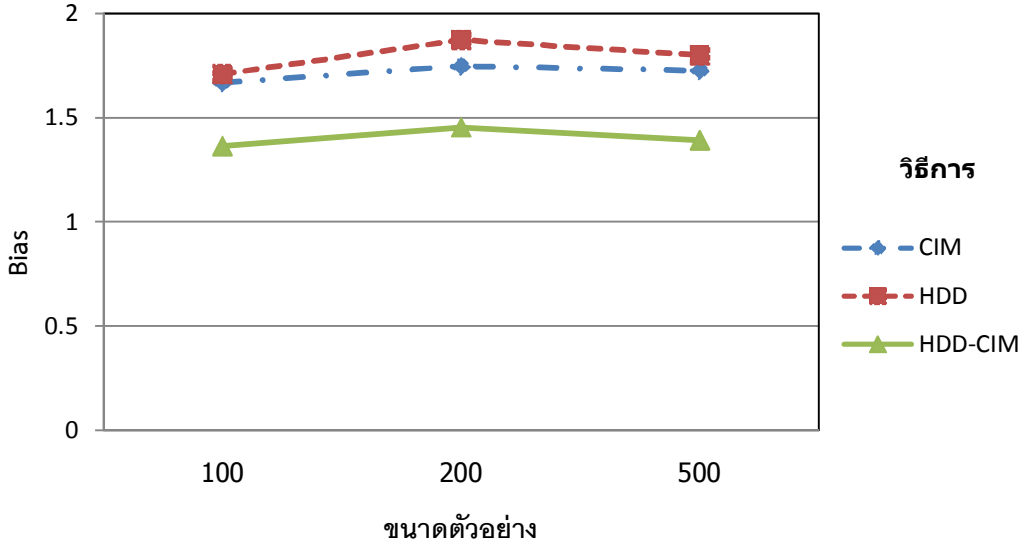
จากตาราง 4.12 พบว่า ไม่มีปฏิสัมพันธ์สามปัจจัยระหว่างขนาดตัวอย่าง ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย นั่นคือการใช้ขนาดตัวอย่างต่างกัน ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ไม่ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

เมื่อพิจารณาปฏิสัมพันธ์สองปัจจัย พบว่า มีปฏิสัมพันธ์สองปัจจัยระหว่าง 1) ขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย และ 2) ร้อยละข้อมูลสูญหายและวิธีการจัดการข้อมูลสูญหาย ที่ระดับนัยสำคัญ .01 นั่นคือ 1) เมื่อใช้ขนาดตัวอย่างต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย 2) เมื่อใช้ร้อยละข้อมูลสูญหายต่างกันและวิธีการจัดการข้อมูลสูญหายต่างกัน ร่วมส่งผลต่อประสิทธิภาพการประมาณค่าความแปรปรวน

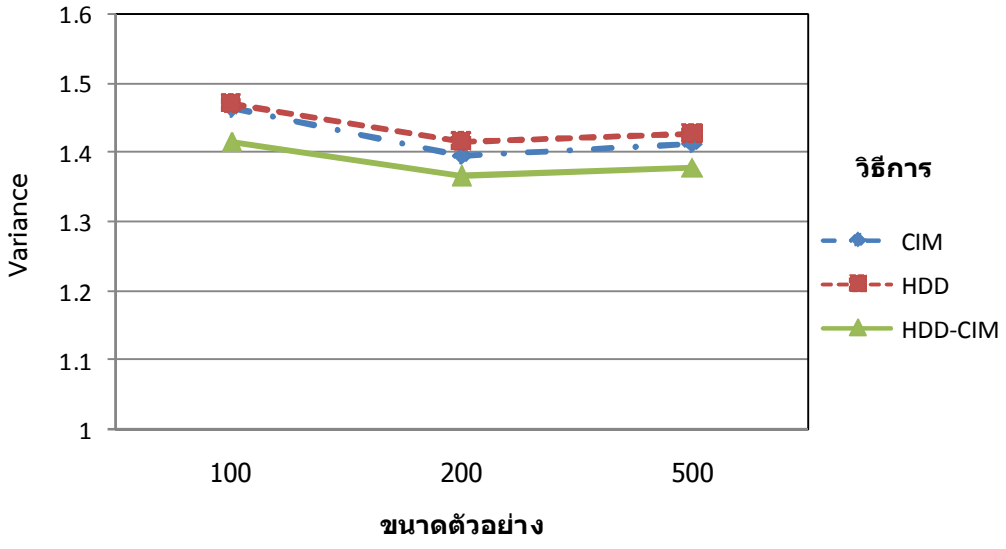
แผนภาพการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียง ความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย แสดงดังภาพประกอบที่ 4.1 – 4.4



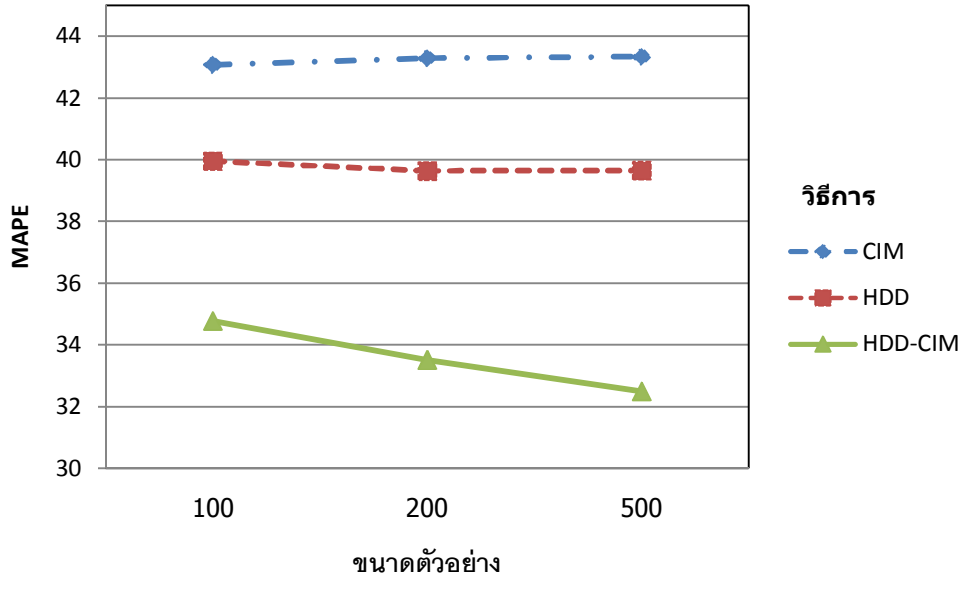
ภาพประกอบ 4.1 เปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย



ภาพประกอบ 4.2 เปรียบเทียบประสิทธิภาพการประมาณค่าความเอนเอียง จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย



ภาพประกอบ 4.3 เปรียบเทียบประสิทธิภาพการประมาณค่าความแปรปรวน จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย



ภาพประกอบ 4.4 เปรียบเทียบประสิทธิภาพค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ จำแนกตามขนาดตัวอย่างและวิธีการจัดการข้อมูลสูญหาย

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM กับวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี CIM และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD ซึ่งเป็นการศึกษาโดยวิธีทดลองที่กระทำกับข้อมูลที่มีอยู่จริง ซึ่งใช้ข้อมูลจากแฟ้มข้อมูลทุติยภูมิจากแฟ้มข้อมูลสำรวจความคิดเห็นของประชาชนเกี่ยวกับสถานการณ์การแพร่ระบาดยาเสพติด พ.ศ. 2554 ที่สำรวจโดยสำนักงานสถิติจังหวัดมหาสารคาม โดยแบ่งขั้นตอนการทำวิจัยเป็น 3 ขั้นตอน ดังนี้

3.1 การพัฒนาวิธีประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM

3.2 การตรวจสอบประสิทธิภาพของวิธีการจัดการข้อมูลสูญหาย

3.3 การวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละของข้อมูลสูญหาย

3.1 การพัฒนาวิธีประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM

การพัฒนาวิธีประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM มีขั้นตอนดังนี้

1. ศึกษาเอกสาร แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้องกับวิธีการประมาณค่าข้อมูลสูญหาย

2. วิเคราะห์แนวคิดของวิธีการประมาณค่าข้อมูลสูญหายแบบต่างๆ โดยพิจารณาข้อดีข้อด้อยของแต่ละวิธี นำข้อดีและข้อด้อยของแต่ละวิธีเพื่อเป็นแนวทางในการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายใหม่ที่จะนำเสนอ

3. นำเสนอวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM

วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM (Hot-Deck-Corrected Item Mean Substitution) ซึ่งมีวิธีการโดยสังเขปต่อไปนี้

วิธี HDD-CIM (Hot-Deck-Corrected Item Mean Substitution) มีหลักการทำงานโดยการแทนที่ข้อมูลที่สูญหายด้วยวิธีที่นำเอาทั้งผลกระทบจากบุคคล (ผู้ตอบ) และผลกระทบจากคำถาม (ตัวแปร) มาร่วมเป็นตัวถ่วงน้ำหนัก เพื่อใช้สร้างข้อมูลทดแทนเพื่อให้มีค่าสังเกตที่สมบูรณ์ครบทุกหน่วยเสียก่อน จากนั้นจึงค่อยคำนวณหาค่าระยะทางระหว่างหน่วยสำรวจแล้วตัดสินใจเลือกหน่วยให้ข้อมูลตามวิธี HDD

ขั้นดำเนินการของวิธี HDD-CIM ปรากฏดังนี้

ขั้นตอนที่ 1 แทนที่รายการข้อมูลที่สูญหายด้วยค่าเฉลี่ยของตัวแปรตามวิธี CIM เพื่อให้ได้แฟ้มข้อมูลสมบูรณ์ชั่วคราว จากสูตร

$$CIM_{vi} = w_v \bar{x}_{.i}^{(i)} = \left(\frac{\sum_{i \in obs(v)} x_{vi}}{\sum_{i \in obs(v)} \bar{x}_{.i}^{(i)}} \right); \; v = 1, 2, \ldots, m_i$$

$$= \frac{\bar{x}_{.i}^{(i)} \times \sum_{i \in obs(v)} x_{vi}}{\sum_{i \in obs(v)} \bar{x}_{.i}^{(i)}} \; ; v = 1, 2, \ldots, m_i$$

โดยที่ $\bar{x}_{.i}^{(i)}$ หมายถึง ค่าเฉลี่ยของตัวแปรที่ i คำนวณมาจากเฉพาะข้อมูลของคนที่ตอบข้อถามข้อที่ i

obs (v) หมายถึง บรรดาคำตอบของผู้ตอบรายที่ v สำหรับคำถามต่าง ๆ เฉพาะข้อที่ตอบ

$i \in obs(v)$ หมายถึง เราสนใจเฉพาะคำตอบของทุกคำถามที่ตอบ คำถามที่ไม่ตอบไม่นำมาพิจารณา แสดงดังตาราง 3.1

ตาราง 3.1 ข้อมูล ค่าเฉลี่ยของตัวแปร และยอดรวมของตัวแปรรวมตามรายบุคคล

คนที่	a1	a2	a3	a4	a5	a6	a7	a8	รวม
1	2	5	3	4	1	1	3	4	**24**
2	1	6	3		1			4	**17**
3		3	3	4		4	2		**19**
4	1	2	7	7	1	4	1	1	**28**
5	2	2	3			2	4	0	**18**
6	2	2		7	1	1			**19**
7		4	5	4	1		3	6	**30**
8	2	4	4	4	3	3	5	8	**41**
9			3	5		4	4		**25**
10	2	1	3	7	1	3	4	3	**34**
รวม	12	29	34	42	9	22	26	26	
เฉลี่ย	**1.71**	**3.22**	**3.78**	**5.25**	**1.29**	**2.75**	**3.25**	**3.71**	**24.96**

ทำการคำนวณค่าจากสูตรข้างต้นเพื่อแทนที่ให้ได้แฟ้มข้อมูลที่สมบูรณ์ชั่วคราว ดังตัวอย่าง

แทนที่ a4 คนที่ 2 (17/24.96)*5.25 = 3.58

แทนที่ a6 คนที่ 2 (17/24.96)*2.75 = 1.87

เมื่อคำนวณจนครบแล้วให้ทำการแทนที่ข้อมูลสูญหายเพื่อให้ได้แฟ้มข้อมูลที่สมบูรณ์ชั่วคราว แสดงดังตาราง 8

ตาราง 3.2 ข้อมูลและข้อมูลทดแทนที่ได้จากขั้นตอนที่ 1

คนที่	a1	a2	a3	a4	a5	a6	a7	a8
1	2	5	3	4	1	1	3	4
2	1	6	3	3.58	1	1.87	2.21	4
3	1.30	3	3	4	0.98	4	2	2.83
4	1	2	7	7	1	4	1	1
5	2	2	3	3.79	0.93	2	4	0
6	2	2	2.88	7	1	1	2.47	2.83
7	2.06	4	5	4	1	3.30	3	6
8	2	4	4	4	3	3	5	8
9	1.72	3.23	3	5	1.29	4	4	3.72
10	2	1	3	7	1	3	4	3

ขั้นตอนที่ 2 คำนวณหาระยะทางระหว่างหน่วยสำรวจ ด้วยวิธีการ HDD สามารถคำนวณได้จากสูตร

$$d_{vv'}^2 = \sum_{i \in obs(v)} (x_{vi} - x_{v'i})^2$$

ขั้นตอนที่ 3 เลือกหน่วยผู้ให้ข้อมูล (หน่วยสำรวจที่สมบูรณ์) โดยแทนที่ข้อมูลที่สูญหายของหน่วยรับข้อมูลด้วยข้อมูลรายการเดียวกันจากหน่วยให้ข้อมูล

ถ้าหน่วยให้ข้อมูลหลายหน่วยมีระยะทางไปยังหน่วยรับข้อมูลสั้นที่สุดแต่มีค่าเท่ากันหลายค่า ให้เลือกหน่วยให้ข้อมูลที่อยู่ในลำดับที่ใกล้หน่วยรับข้อมูลกว่าเป็นหน่วยให้ข้อมูล

ตาราง 3.3 ค่าระยะทางระหว่างหน่วยสำรวจตามขั้นตอนที่ 2

	1	2	3	4	5	6	7	8	9	10
1	0	1.887117	3.982467	7.247449	5.201093	4.434965	3.783563	5.477226	3.792845	5.567764
2	1.887117	0	3.901406	7.663365	6.021507	5.563361	3.996772	5.934464	4.256067	6.572179
3	3.982467	3.901406	0	5.516418	4.186102	4.441135	4.138	6.582607	2.472574	4.302771
4	7.549834	7.663365	5.516418	0	6.429536	5.702865	6.899797	9.643651	6.204655	5.656854
5	5.201093	6.021507	4.186102	6.429536	0	4.655742	6.83793	8.680047	4.585409	4.619408
6	4.434965	5.563361	4.441135	5.702865	4.655742	0	5.759844	7.443734	4.220264	2.715812
7	3.783563	3.996772	4.138	6.899797	6.83793	5.759844	0	3.618904	3.532988	5.665374
8	5.477226	5.934464	6.582607	9.643651	8.680047	7.443734	3.618904	0	5.092379	7
9	3.792845	4.256067	2.472574	6.204655	4.585409	4.220264	3.532988	5.092379	0	3.261829

ทำการเลือกหน่วยผู้ให้ข้อมูล (หน่วยสำรวจที่สมบูรณ์) โดยแทนที่ข้อมูลที่สูญหายของหน่วยรับข้อมูลด้วยข้อมูลรายการเดียวกันจากหน่วยให้ข้อมูล แสดงดังตาราง 3.4

ตาราง 3.4 ข้อมูลและข้อมูลทดแทนตามขั้นตอนที่ 3

คนที่	a1	a2	a3	a4	a5	a6	a7	a8
1	2	5	3	4	1	1	3	4
2	1	6	3	**4.00**	1	**1.00**	**3.00**	4
3	**2.00**	3	3	4	**1.00**	4	2	**4.00**
4	1	2	7	7	1	4	1	1
5	2	2	3	**7.00**	**1.00**	2	4	0
6	2	2	**3.00**	7	1	1	**4.00**	**3.00**
7	**2.00**	4	5	4	1	**3.00**	3	6
8	2	4	4	4	3	3	5	8
9	**2.00**	**1.00**	3	5	**1.00**	4	4	**3.00**
10	2	1	3	7	1	3	4	3

3.2 การตรวจสอบประสิทธิภาพของวิธีการจัดการข้อมูลสูญหาย

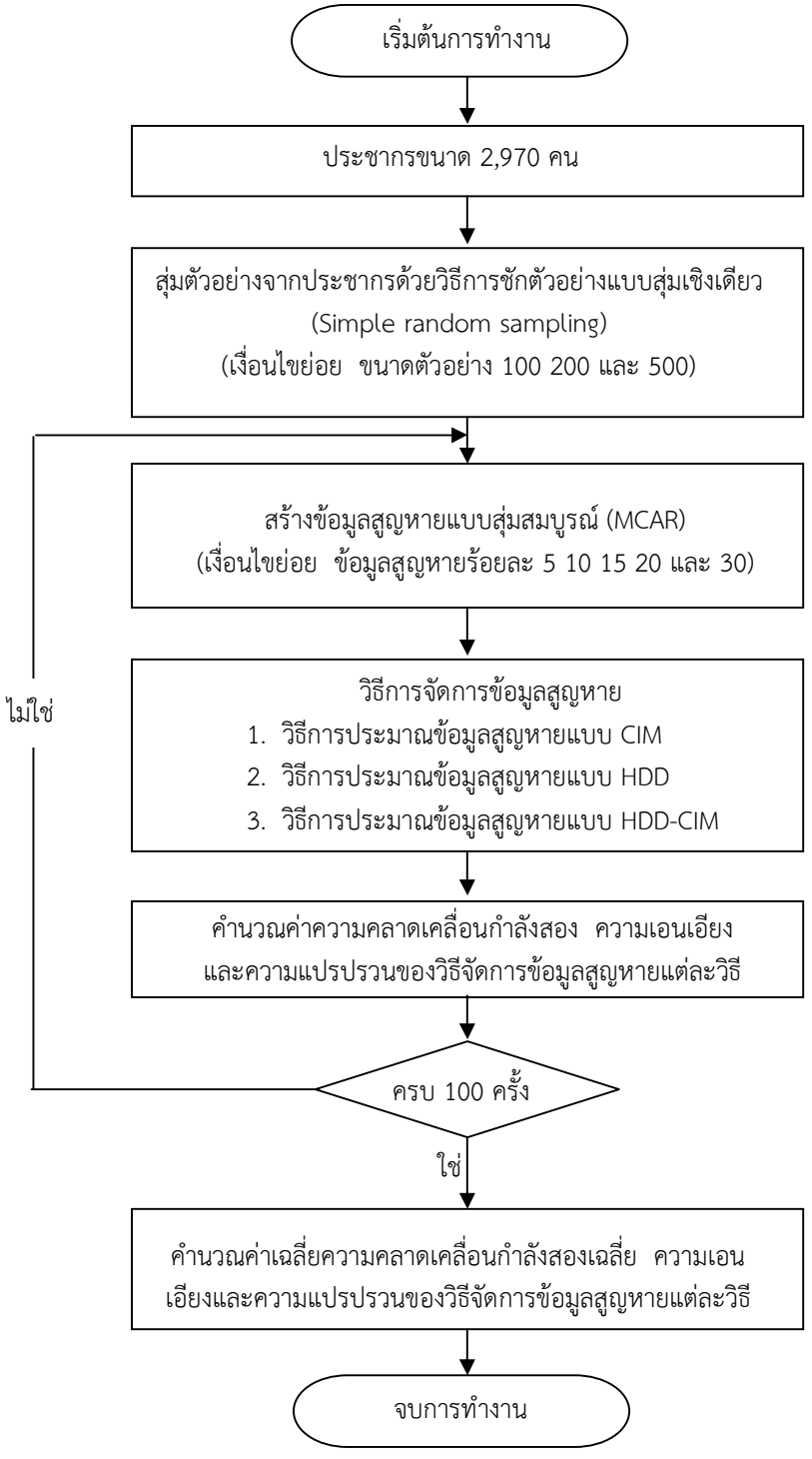
การตรวจสอบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM กับวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี CIM และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD มีรายละเอียดดังนี้

1. สุ่มตัวอย่างจากประชากรด้วยวิธีการชักตัวอย่างแบบสุ่มเชิงเดียว (Simple Random Sampling) โดยมีขนาดตัวอย่างตามที่กำหนด
2. สร้างข้อมูลสูญหายแบบสุ่มสมบูรณ์ (MCAR) โดยมีร้อยละของข้อมูลสูญหายตามที่กำหนด
3. จัดการข้อมูลสูญหายด้วยวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี CIM และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD
4. คำนวณค่าความคลาดเคลื่อนกำลังสอง ค่าความเอนเอียง และค่าความแปรปรวนของวิธีการจัดการข้อมูลสูญหายของแต่ละวิธี
5. ทำซ้ำจากขั้นตอนที่ 2 ถึงขั้นตอนที่ 4 จำนวน 100 ซ้ำ
6. คำนวณค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง ค่าความเอนเอียง ของวิธีการจัดการข้อมูลสูญหายของแต่ละวิธี

ดำเนินการในลักษณะเดียวกันจากขั้นตอนที่ 2 ถึงขั้นตอนที่ 6 โดยกำหนดขนาดตัวอย่างและร้อยละของข้อมูลสูญหาย ดังต่อไปนี้

1. ขนาดตัวอย่าง 100 200 และ 500
2. ร้อยละของข้อมูลสูญหายร้อยละ 5 10 15 20 และ 30

การดำเนินการเปรียบเทียบประสิทธิภาพวิธีการจัดการข้อมูลสูญหายสรุปเป็นแผนภาพดังภาพประกอบ 1



ภาพประกอบ 3.1 ขั้นตอนการเปรียบเทียบประสิทธิภาพวิธีการจัดการข้อมูลสูญหาย

3.3 การเปรียบเทียบประสิทธิภาพระหว่างวิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละของข้อมูลสูญหาย

การเปรียบเทียบประสิทธิภาพระหว่างวิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละของข้อมูลสูญหาย ที่มีต่อประสิทธิภาพการประมาณค่าเฉลี่ยความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ จากการศึกษางานวิจัยพบว่า มีปฏิสัมพันธ์ระหว่างวิธีประมาณค่าข้อมูลสูญหาย ระดับความสัมพันธ์ของตัวแปรและร้อยละการสูญหายของข้อมูล (จำลอง วงษ์ประเสริฐ, 2554: 100 ; อ้างอิงจาก Enders, 2001: 713-740) และเชาว์ อินใย (2552) พัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบอีพีอาร์และตรวจสอบความแม่นยำ และอำนาจการทดสอบที่ได้จากวิธีการประมาณค่าข้อมูลสูญหายแบบอีพีเอสเอสอี วิธีการประมาณค่าข้อมูญสูญหายแบบอีเอ็มและการตัดข้อมูลแบบลิสไวส์ ผลการศึกษาพบว่า ไม่มีปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย ร้อยละข้อมูลสูญหายและระดับความสัมพันธ์ระหว่างตัวแปร ที่มีผลต่อความแม่นยำของการประมาณค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ แต่พบปฏิสัมพันธ์ระหว่างวิธีการจัดการข้อมูลสูญหาย ร้อยละข้อมูลสูญหายและระดับความสัมพันธ์ระหว่างตัวแปร ที่มีผลต่อความแม่นยำของการประมาณค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาปฏิสัมพันธ์ระหว่างวิธีการจัดการข้อมูลสูญหายได้แก่ วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM กับวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี CIM และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD กับร้อยละของข้อมูลสูญหายว่ามีผลต่อประสิทธิภาพในการประมาณค่าความคลาดเคลื่อนกำลังสอง ค่าความเอนเอียงและค่าความแปรปรวนหรือไม่ โดยการทำการหาปฏิสัมพันธ์ทีละตัวแปร โดยใช้วิธีการวิเคราะห์ความแปรปรวน 3 ทาง (Three – way Analysis of Variance : 3 – way ANOVA)

ขั้นตอนการวิเคราะห์ความแปรปรวน มีดังนี้

1. ตรวจสอบข้อตกลงเบื้องต้นในการวิเคราะห์ความแปรปรวน โดยข้อตกลงเบื้องต้นที่ต้องทำการทดสอบได้แก่ ตัวอย่างเป็นอิสระกัน ความแปรปรวนของตัวแปรตามในแต่ละกลุ่มเท่ากัน และตัวแปรตามมีการแจกแจงแบบปกติ

2. วิเคราะห์ความแปรปรวน โดยใช้สถิติ F คำนวณจากสูตรดังนี้

$$SS_A = bcn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y})^2$$

$$SS_B = acn\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y})^2$$

$$SS_C = abn\sum_{k=1}^{c}(\bar{y}_{..k} - \bar{y})^2$$

$$SS_{AB} = cn\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y})^2$$

$$SS_{AC} = bn\sum_{i=1}^{a}\sum_{k=1}^{c}(\bar{y}_{i.k} - \bar{y}_{i..} - \bar{y}_{..k} - \bar{y})^2$$

$$SS_{BC} = an\sum_{j=1}^{b}\sum_{k=1}^{c}(\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} - \bar{y})^2$$

$$SS_{ABC} = an\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\bar{y}_{ijk} - \bar{y}_{ij.} - \bar{y}_{i.k} - \bar{y}_{.jk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} - \bar{y})^2$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}(\bar{y}_{ijkm} - \bar{y})^2$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}(y_{ijkm} - \bar{y})^2$$

ตาราง 3.5 การวิเคราะห์ความแปรปรวนแบบสามทาง (Three – way Analysis of Variance)

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
A	SS_A	*a – 1*	$MS_A = \frac{SS_A}{(a-1)}$	$F_A = \frac{MS_A}{MS_E}$
B	SS_B	*b – 1*	$MS_B = \frac{SS_B}{(b-1)}$	$F_B = \frac{MS_B}{MS_E}$
C	SS_C	*c – 1*	$MS_C = \frac{SS_C}{(c-1)}$	$F_C = \frac{MS_C}{MS_E}$
AB	SS_{AB}	*(a – 1)(b – 1)*	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{MS_{AB}}{MS_E}$
AC	SS_{AC}	*(a – 1)(c – 1)*	$MS_{AC} = \frac{SS_{AC}}{(a-1)(c-1)}$	$F_{AC} = \frac{MS_{AC}}{MS_E}$
BC	SS_{BC}	*(b – 1)(c – 1)*	$MS_{BC} = \frac{SS_{BC}}{(b-1)(c-1)}$	$F_{BC} = \frac{MS_{BC}}{MS_E}$
ABC	SS_{ABC}	*(a – 1)(b – 1)(c – 1)*	$MS_{ABC} = \frac{SS_{ABC}}{(a-1)(b-1)(c-1)}$	$F_{ABC} = \frac{MS_{ABC}}{MS_E}$
Error	SS_E	*abc (n – 1)*	$MS_E = \frac{SS_E}{abc(n-1)}$	
Total	SS_T	*abcn – 1*		

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การประมาณค่าข้อมูลสูญหาย ผู้ศึกษาค้นคว้าได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องโดยนำเสนอเนื้อหาตามลำดับดังนี้

2.1 ข้อมูลสูญหาย

2.2 การประมาณค่าข้อมูลสูญหายแบบ Hot-Deck

2.3 การประมาณค่าข้อมูลสูญหายแบบ Corrected Item Mean Substitution (CIM)

2.4 งานวิจัยที่เกี่ยวข้อง

2.4.1 งานวิจัยในประเทศ

2.4.2 งานวิจัยต่างประเทศ

2.1 ข้อมูลสูญหาย

2.1.1 ความหมายของข้อมูลสูญหาย

ข้อมูลสูญหายเป็นกรณีที่พบได้บ่อยในงานวิจัยทุกสาขา ซึ่งนักวิจัยจำเป็นต้องพิจารณาถึงวิธีการที่เหมาะสมสำหรับใช้ในการจัดการข้อมูลสูญหายในทุก ๆ ครั้งที่พบกับปัญหานี้ วิธีการที่ใช้สำหรับจัดการกับข้อมูลสูญหายมีวิธีการให้พิจารณาเลือกใช้ค่อนข้างมาก หากเลือกใช้วิธีการจัดการข้อมูลสูญหายที่ไม่เหมาะสมกับสถานการณ์ของข้อมูลสูญหาย ย่อมส่งผลทำให้เกิดการบิดเบือนผลการวิเคราะห์และการสรุปผลการวิจัย

ข้อมูลสูญหาย คือ ค่าสังเกตที่ต้องการทราบค่าแต่ไม่สามารถทราบค่าได้ โดยที่ค่านั้นควรจะสามารถทราบค่าได้หากวิธีการที่ใช้ในการรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีขึ้นหรือมีความเหมาะสมมากขึ้น ข้อมูลสูญหายเป็นปัญหาที่พบโดยทั่วไปในการวิจัยเชิงปริมาณ เป็นเรื่องที่ต้องเกิดขึ้นที่นักวิจัยจะต้องเผชิญ แม้ว่าจะได้ควบคุมการสำรวจหรือการทดลองไว้แล้วเป็นอย่างดีก็ตาม (Huisman, 1998: 271-278) นักวิจัยได้ให้ความสำคัญและให้ความสำคัญเพิ่มขึ้นเรื่อย ๆ (Carlson, 2001) วิธีการทางสถิติได้พัฒนาเพื่อการวิเคราะห์ข้อมูลสำหรับชุดข้อมูลที่สมบูรณ์ แต่เมื่อมีข้อมูลสูญหายย่อมส่งผลกระทบต่อประสิทธิภาพการวิเคราะห์ข้อมูล ในอดีตข้อมูลสูญหายได้ถูกจัดการโดยวิธีการต่าง ๆ วิธีการที่พยายามแก้ไขข้อมูลก่อนการวิเคราะห์ เช่น การลบข้อมูลที่ไม่สมบูรณ์ทิ้งเป็นหนึ่งในกลยุทธ์ดังกล่าว วิธีอื่น ๆ เช่น การแทนค่าตัวแปรที่หายไปโดยใช้ค่าเฉลี่ยของตัวแปรที่เหลืออยู่ แต่วิธีการเหล่านี้ทำให้ตัวประมาณเกิดความเอนเอียง

นักวิจัยพยายามจะขจัดปัญหาข้อมูลสูญหาย โดยพยายามที่จะเก็บข้อมูลให้ได้ข้อมูลที่สมบูรณ์ แต่ปัญหาข้อมูลสูญหายนั้นเกิดจากหลากหลายสาเหตุ ตัวอย่างเช่น ในการวิจัยเชิงสำรวจข้อมูลที่ไม่สมบูรณ์ เกิดจากหน่วยตัวอย่างไม่ตอบสนองหรือการไม่ตอบสนองบางข้อคำถาม ในการศึกษาระยะยาว ข้อมูลอาจสูญหายเนื่องจากความผิดพลาดเกี่ยวกับนัดหมาย การเจ็บป่วยของหน่วยที่ให้ข้อมูล การขาดแรงจูงใจในการให้ข้อมูล หรือเกิดจากการย้ายถิ่นฐานของหน่วยที่ให้ข้อมูลทางด้านการศึกษานั้น อาจเกิดจากการขาดสอบของนักเรียน การปฏิเสธการให้ข้อมูลบางประการ

ปัญหาข้อมูลสูญหายอาจถือว่าเป็นปัญหาที่ไม่รุนแรง ถ้าการวิเคราะห์ข้อมูลกระทำด้วยสถิติที่วิเคราะห์ข้อมูลตัวแปรเดี่ยว เช่น ร้อยละ ค่าเฉลี่ยหรือสถิติพรรณนาอื่น แต่ถ้าการวิเคราะห์ข้อมูลนั้นจำเป็นต้องใช้วิธีวิเคราะห์ข้อมูลตัวแปรพหุ เช่น การวิเคราะห์ความถดถอยพหุ การวิเคราะห์เส้นทาง การวิเคราะห์ปัจจัย การวิเคราะห์กลุ่ม การวิเคราะห์การจำแนกและการถดถอยโลจีสติกส์ เป็นต้น ในกรณีนี้การสูญหายของข้อมูลจะมีผลกระทบที่รุนแรง เพราะถ้าพบว่าหน่วยวิเคราะห์ใดมีตัวแปรใดขาดข้อมูลไปแม้เพียงตัวแปรเดียว ก็จะตัดหน่วยวิเคราะห์นั้นทิ้งทั้งหน่วย โดยไม่สนใจว่าจะยังมีตัวแปรอื่นอีกมากที่มีข้อมูลครบถ้วนหรือไม่ (Heeringa, 2000; Roth, 1994: 537 – 560)

การตัดทิ้งข้อมูลเป็นวิธีปริยายของการวิเคราะห์ข้อมูลตัวแปรพหุในโปรแกรม SPSS หรือโปรแกรมสถิติทั่วไป (Kim and Curry, 1977: 215 – 240) ได้ทำการศึกษาโดยวิธีทดลองพบว่า ถ้าตัวแปรแต่ละตัวมีข้อมูลหายไปโดยสุ่มเพียงร้อยละ 10 จะมีผลให้ต้องตัดหน่วยวิเคราะห์ทิ้งถึงร้อยละ 59 เห็นได้ว่าเป็นความสูญเสียในอัตราที่สูงมาก การวิเคราะห์ข้อมูลจากเฉพาะข้อมูลที่เหลืออยู่ภายหลังจากที่ได้ตัดทิ้งค่าสังเกตที่ไม่สมบูรณ์ไปแล้ว ผลการวิเคราะห์จะเอนเอียงและไม่ถูกต้อง (จำลอง วงษ์ประเสริฐ, 2554: 18; อ้างอิงจาก Wang, 2000)

การตัดหน่วยสำรวจที่ข้อมูลไม่สมบูรณ์ทิ้งเป็นการสูญเสียสารสนเทศ เสียค่าใช้จ่าย ดังนั้นควรหาวิธีการในการประมาณค่าข้อมูลที่สูญหายไปนั้น ด้วยวิธีที่เหมาะสมก่อนทำการวิเคราะห์ข้อมูล เพราะหากได้มีการทดแทนรายการข้อมูลที่สูญหายด้วยวิธีการที่เหมาะสมก่อนการวิเคราะห์ข้อมูลจะพบว่า สถิติทดสอบและสถิติอื่น ๆ มีอำนาจทดสอบสูงขึ้น (Wu, 1989)

จึงกล่าวได้ว่า ข้อมูลสูญหาย คือ ค่าสังเกตที่ต้องการทราบแต่ไม่สามารถทราบค่าได้ โดยที่ค่านั้นควรจะสามารถทราบค่าได้ หากวิธีการที่ใช้ในการรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีขึ้น หรือมีความเหมาะสมมากขึ้น

การวิจัยโดยวิธีสำรวจไม่ว่าจะสำรวจโดยวิธีใด เช่นโดยวิธีสัมภาษณ์ซึ่งเป็นการพบกันโดยตรงระหว่างผู้ให้ข้อมูลกับพนักงานสำรวจผู้ขอสัมภาษณ์ (Face-to-face Interview) หรือการสัมภาษณ์ทางโทรศัพท์ซึ่งเป็นการสัมภาษณ์แบบไม่เห็นตัวและไม่เห็นหน้าตากันได้ยินเฉพาะเสียงและสำเนียงภาษา หรือการสอบถามโดยทางไปรษณีย์ และการสอบถามผ่านอินเตอร์เน็ต ล้วนแต่เป็นเรื่องที่เกี่ยวข้องกับคนและสังคม การจะได้รับข้อมูลจากผู้ตอบหรือไม่ ครบถ้วนเพียงใดจึงเป็นเรื่องที่เกี่ยวข้องกับทฤษฎีทางสังคมหลายทฤษฎีที่มีผลร่วมกัน ซึ่งอาจประกอบไปด้วยทฤษฎีปริวรรตนิยม (Social Exchange Theory) ทฤษฎีปฏิสัมพันธ์สัญลักษณ์ (Symbolic Interaction Theory) และทฤษฎีการโดดเดี่ยวทางสังคม (Social Isolation Theory หรือ Social Location Theory) (มนตรี พิริยะกุล, 2548: 15-18)

2.1.2 การสูญหายของข้อมูล

การสูญหายของข้อมูลเป็นเหตุการณ์ปกติที่เกิดขึ้นได้เสมอในงานวิจัยทั่วไป และอาจรุนแรงมากขึ้นในงานวิจัยโดยวิธีทดลอง (Lepkowski and Stepouwer, 1987; อ้างอิงจาก Roth, 1995) และในงานวิจัยโดยวิธีสำรวจ (Kim and Curry, 1977; อ้างอิงจาก Roth, 1995) สาเหตุหลักของข้อมูลสูญหายคือการไม่ตอบสนองของหน่วยสำรวจ คำว่าการไม่ตอบสนองของหน่วยสำรวจ (Nonresponse) เป็นคำรวม หมายถึงผู้ตอบไม่ให้ความร่วมมือโดยสิ้นเชิง (Unit Nonresponse) กับผู้ตอบยินดีตอบบางข้อและปฏิเสธที่จะตอบในบางข้อ (Item Non Response) การศึกษาครั้งนี้จะหมายถึงกรณีหลังนี้เท่านั้น

การไม่ตอบแบบสำรวจบางข้อคือกรณีที่ผู้ตอบยินยอมให้ความร่วมมือในการตอบแบบสำรวจบางส่วนและไม่ร่วมมือบางส่วน การไม่ตอบจึงอาจเกิดขึ้นกับข้อถามข้อใดก็ได้ไม่อาจกำหนดที่อยู่ได้แน่นอน จากการวิจัยพบว่าข้อถามข้อแรก ๆ คือ ส่วนต้นของแบบสำรวจมักมีปัญหาการไม่ตอบน้อยกว่าข้อถามในส่วนกลาง ๆ หรือท้าย ๆ ดังนั้นเราอาจแยกว่าการสูญหายของข้อมูลที่เกิดจากการไม่ตอบแบบสำรวจตั้งแต่ข้อต้น ๆ คือ การไม่ตอบโดยสิ้นเชิง (Unit Nonresponse) และเรียกการสูญหายของข้อมูลที่เกิดขึ้นในส่วนท้าย ๆ ของแบบสำรวจว่าการไม่ตอบบางข้อ (Item Nonresponse) แต่มิได้หมายความว่าการไม่ตอบแบบสำรวจจะหมายถึงการไม่ตอบเฉพาะข้อท้าย ๆ เรื่องนี้มีเหตุผลอื่น ๆ ประกอบอีกหลายประการถึงสาเหตุของการไม่ตอบบางข้อ

การไม่ตอบแบบสำรวจเป็นปัญหาสำคัญของงานวิจัยประเภทสำรวจ มีผลกระทบให้เกิดการสูญหายข้อมูลบางตัวแปร ทำให้สูญเสียสารสนเทศเพราะต้องตัดหน่วยสำรวจนั้นทิ้งไป กลุ่มตัวอย่างที่เหลืออยู่จึงน้อยลงกว่าที่กำหนดเอาไว้เดิม กลุ่มตัวอย่างที่เหลือจึงมิใช่ตัวแทนที่ดีหรือมีความบกพร่องในการเป็นตัวแทนของประชากร การวิเคราะห์ข้อมูลจากตัวอย่างเท่าที่เหลืออยู่จึงสูญเสียสภาพสุ่มไปบางส่วน ทั้งมีจำนวนที่ต่ำกว่าที่กำหนด จึงมีความเป็นไปได้ที่ผลการวิเคราะห์จะนำสู่ความเอนเอียง (Bias) แม้ว่าจะใช้เครื่องมือการวิเคราะห์ข้อมูลมาตรฐานที่พิสูจน์แล้วว่าให้ค่าประมาณที่ไม่เอนเอียงก็ตาม

2.1.3 สาเหตุสำคัญของการเกิดข้อมูลสูญหาย (Major Reason of Missing Data)

ในการเกิดข้อมูลสูญหายมักเป็นผลอันเนื่องมาจากหลายกรณี โดยเหตุผลพื้นฐานมักเป็นผลจาก (ปิยะภรณ์ ประสิทธิ์วัฒนาเสรี และสุคนธ์ ประสิทธิ์วัฒนาเสรี, 2552: 52 - 61)

2.1.3.1 การไม่ยอมมาแสดงตัวของหน่วยตัวอย่างในช่วงเวลาเฝ้าติดตามหน่วยตัวอย่าง

2.1.3.2 หน่วยตัวอย่างปฏิเสธการตอบคำถามในบางคำถามของแบบฟอร์มที่ใช้รวบรวมข้อมูล ซึ่งส่วนใหญ่มักเป็นคำถามที่กระทบต่อความรู้สึกได้ง่าย

2.1.3.3 หน่วยตัวอย่างไม่ทราบคำตอบ ทั้งนี้อาจเป็นผลมาจากปัญหาในเรื่องของความจำ

2.1.3.4 คำถามที่ใช้ไม่ครอบคลุมทุกกรณีจึงทำให้เกิดข้อมูลสูญหาย

2.1.3.5 การนำข้อมูลเข้าสู่ระบบประมวลผลทางคอมพิวเตอร์ ซึ่งอาจเป็นผลจากความผิดพลาดของระบบฐานข้อมูล การเชื่อมโยงระหว่างข้อมูลจากฝ่ายต่าง ๆ

2.1.4 การลดขนาดของข้อมูลสูญหาย (Minimize Missing Data)

แนวทางในการลดผลกระทบที่เกิดขึ้นจากข้อมูลสูญหายที่ดีที่สุดคือ การป้องกันให้ไม่เกิดข้อมูลสูญหายในงานวิจัยหรือหากจะเกิดข้อมูลสูญหายขึ้นต้องพยายามลดขนาดของจำนวนข้อมูลสูญหายให้เหลือจำนวนน้อยที่สุด โดยสิ่งสำคัญในการหาทางป้องกันหรือลดขนาดของข้อมูลสูญหายก็คือความเข้าใจถึงสาเหตุของการเกิดข้อมูลสูญหายนั้น ๆ เพื่อจะได้ทำการพัฒนาโครงสร้างแผนการศึกษาที่สอดคล้องกับวัตถุประสงค์ในการศึกษา มีความครอบคลุมชัดเจน และสามารถตอบปัญหาในทุกประเด็นที่ศึกษาได้อย่างสมบูรณ์ นอกจากนี้ในระหว่างขั้นตอนการเก็บรวบรวมข้อมูลทางโทรศัพท์หรือจากการสัมภาษณ์ หรือจากการตรวจวัด ควรต้องมีการตรวจเช็คความสมบูรณ์ของข้อมูลก่อนที่จะดำเนินการเสร็จสิ้น

แนวทางหนึ่งสำหรับลดการปฏิเสธที่จะตอบข้อคำถามจากหน่วยตัวอย่างคือ การพยายามพัฒนาตัวแบบสอบถามให้ดีขึ้นและรัดกุม พร้อมทั้งการอธิบายถึงวัตถุประสงค์ของการศึกษา รวมถึงประโยชน์ที่จะเกิดขึ้นของผลการศึกษาในครั้งนี้ต่อหน่วยตัวอย่าง นอกจากนี้ควรต้องมีการรับรองในเรื่องข้อมูลที่ให้มาจะถูกเก็บเป็นความลับเสมอ ควรลดลักษณะของคำถามที่กระทบต่อความรู้สึกได้ง่ายหรืออาจเลี่ยงใช้คำถามในลักษณะที่ไม่ใช่การถามตรง (Sudman, 1982)

สำหรับแนวทางในการลดคำตอบที่ว่า “ไม่ทราบ” สามารถทำได้โดยเพิ่มเติมส่วนขยายความของข้อมูลที่ต้องการหรือให้แนวทางของคำตอบ หรือเทคนิคเพื่อช่วยในการหาคำตอบแก่หน่วยตัวอย่าง ยกตัวอย่างเช่น แทนที่จะถามว่า “ท่านทราบโทษภัยของยาเสพย์ติดหรือไม่” อาจทำการจัดหารายการหรือกิจกรรมที่เกี่ยวข้องกับโทษภัยของยาเสพย์ติดมาให้ จากนั้นให้ทำการตอบว่าทราบ/ไม่ทราบ ข้างท้ายแต่ละรายการ ลักษณะของคำถามที่ใช้นี้นอกจากจะช่วยในการให้คำนิยามเกี่ยวกับโทษภัยของยาเสพย์ติดแล้ว ยังเป็นการให้แนวทางของคำตอบที่ช่วยในการหาคำตอบแก่หน่วยตัวอย่างอีกด้วย ในส่วนของข้อมูลสูญหายที่เกิดขึ้นเนื่องจากความผิดพลาดในขั้นตอนการดำเนินการกับข้อมูล (Data Processing) สามารถลดความผิดพลาดนี้ได้โดยการออกแบบแบบสอบถามที่สะดวกต่อการนำเข้าข้อมูลรวมถึงการพัฒนาหรือเลือกวิธีการที่ใช้ในการนำเข้าข้อมูลให้เหมาะสม (O’Rourke, 2000)

2.1.5 การตรวจเช็คข้อมูลสูญหาย (Detect Missing Data)

สำหรับวิธีการตรวจเช็คข้อมูลสูญหายสามารถทำได้หลายวิธีด้วยกัน (O’Rourke. 2000) ได้อธิบายต่าง ๆ ที่สามารถใช้ในการตรวจเช็คข้อมูล ดังนี้

2.1.5.1 การตรวจเช็คด้วยสายตา (Visual Scanning)

2.1.5.2 โดยใช้โปรแกรมนำเข้าข้อมูล (Data Entry Program) เช่น QPL หรือ SPSS ช่วยในการตรวจเช็ค

2.1.5.3 โดยใช้วิธีการแจกแจงความถี่ของคำตอบในตัวแปรแต่ละตัว

2.1.5.4 สำหรับในการวิเคราะห์ตัวแปรคู่ (Bivariate Analysis) ใช้วิธีการสร้างตารางไขว้ (Crosstabulation) ระหว่างตัวแปรทั้งคู่

2.1.6 ประเภทของข้อมูลสูญหาย (Type of Missing Data)

การพิจารณาประเภทของข้อมูลสูญหายเป็นขั้นตอนที่สำคัญ ทั้งนี้เพราะหากสามารถทราบถึงลักษณะของข้อมูลสูญหายจะช่วยในการพิจารณาแนวทางสำหรับจัดการกับปัญหาความไม่สมบูรณ์ของข้อมูลได้อย่างเหมาะสม ซึ่งโดยทั่วไปมักจำแนกข้อมูลสูญหายออกเป็น 3 ประเภทด้วยกัน คือ (Little and Rubin, 2002)

2.1.6.1 การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random : MCAR)

การสูญหายแบบสุ่มอย่างสมบูรณ์ เป็นลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด นั่นคือข้อมูลที่สูญหายเป็นอิสระจากตัวแปรต่าง ๆ สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหายกลุ่มนี้โดยการแบ่งกลุ่มของค่าสังเกตเป็นกลุ่มข้อมูลปกติและข้อมูลสูญหาย ในกรณีนี้เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างทั้งสองกลุ่มสำหรับตัวแปรต่าง ๆ ในฐานข้อมูลสำหรับสาเหตุที่ทำให้ข้อมูลเกิดการสูญหายมีอยู่หลากหลายเหตุผล อาจเกิดขึ้น

เนื่องจากเครื่องมือเสีย อุปกรณ์เกิดข้อบกพร่อง สภาพอากาศเลวร้าย กลุ่มเป้าหมายที่ศึกษาล้มป่วย หรือการนำเข้าข้อมูลไม่ถูกต้อง

สำหรับข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นสามารถเลือกทำการวิเคราะห์ข้อมูลในส่วนที่สมบูรณ์ได้

2.1.6.2 การสูญหายแบบสุ่ม (Missing at Random : MAR)

การสูญหายแบบสุ่ม เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือ ค่าของข้อมูลสูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในฐานข้อมูลซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย ยกตัวอย่างเช่น หากพบว่าเฉพาะกลุ่มผู้ได้รับการศึกษาน้อยที่ไม่ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการเสพยาเสพย์ติด ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการเสพยาเสพย์ติดมีค่าสูญหายแบบ MAR ทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะในบางส่วนของตัวแปรระดับการศึกษาสำหรับข้อมูลสูญหายประเภทนี้ยังไม่ส่งผลกระทบรุนแรงเท่ากับข้อมูลสูญหายในประเภทสุดท้าย

2.1.6.3 การสูญหายแบบไม่สุ่ม (Not Missing at Random : NMAR)

การสูญหายแบบไม่สุ่ม เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรตัวอื่นด้วย เช่น หากข้อมูลสูญหายของระดับรายได้ขึ้นอยู่ระดับรายได้ในแต่ละช่วงอายุ ข้อมูลสูญหายที่เกิดขึ้นจัดอยู่ในประเภท NMAR หรือในบางกรณีค่าของอมูลสูญหายอาจไม่ขึ้นอยู่กับตัวแปรใด ๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้น แต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกฐานข้อมูล

ลักษณะข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลสูญหายที่สามารถส่งผลกระทบอย่างรุนแรงในการวิเคราะห์ข้อมูล

2.1.7 ความเอนเอียง (Biasness)

ความเอนเอียง (Biasness) หมายถึง สถานการณ์ที่ค่าประมาณมิได้ชี้ไปที่ค่าจริงของสิ่งที่ต้องการศึกษา เป็นการมองในลักษณะเฉลี่ย ค่าประมาณอาจชี้ไปที่ใดก็ได้ ต่ำกว่าค่าจริงก็ได้สูงกว่าค่าจริงก็ได้ นักวิจัยจึงสรุปผลการวิจัยผิดพลาดโดยไม่รู้ตัว

การไม่ตอบแบบสอบถามทั้งชนิดไม่ตอบโดยสิ้นเชิงและชนิดไม่ตอบบางส่วนเป็นเรื่องธรรมชาติเดิมของสังคม เป็นสิ่งที่ต้องเป็นเช่นนี้ (Krohn, 1992) หมายความว่า มนุษย์ย่อมต้องการการแลกเปลี่ยนที่ยุติธรรมหรือต่างก็ได้ประโยชน์เต็มที่ด้วยกันทั้งสองฝ่าย การได้ผลตอบแทนที่ไม่ยุติธรรมที่เป็นวัตถุหรือด้านจิตใจจะทำให้ไม่ร่วมมือ เป็นเรื่องที่เป็นไปตามทฤษฎีแลกเปลี่ยนทางสังคม (Murata, 2001) ในขณะเดียวกันสัญลักษณ์ที่ใช้สื่อสารต้องมีความเหมาะสม เข้าใจได้ตรงกันทั้งสองฝ่ายในบริบทที่กำหนด และพบว่าปฏิสัมพันธ์ตามบริบท (Interaction Context) มีผลต่อการสูญหายของข้อมูลมากกว่าคุณลักษณะของผู้ตอบเสียอีก

การไม่ตอบข้อถามเป็นเรื่องพื้นฐานทางสังคม จิตวิทยาสังคม หรือบุคลิกภาพ เป็นเรื่องการสื่อสารติดต่อกันผ่านสัญลักษณ์ เช่น ใช้ภาษาที่ฟังง่าย เข้าใจเรื่องที่จะถามได้ มีการสื่อสารให้ถามต่อไปหรือย้ายข้ามไปตอบข้ออื่น ๆ ที่ชัดเจน พนักงานสัมภาษณ์จึงต้องเป็นผู้ที่มีบุคลิกลักษณะที่น่า

ไว้ใจ เป็นมิตรและมีมารยาท เหตุนี้กระบวนการทางสถิติในเรื่องการแก้ไขปัญหาข้อมูลสูญหายจึงให้เริ่มที่การให้ความสนใจสาเหตุทางสังคมโดยควบคุมสาเหตุการไม่ตอบให้ได้มากที่สุดเสียก่อน

อย่างไรก็ตาม ต้องไม่ลืมว่าคุณลักษณะทางประชากรรวมถึงบุคลิกภาพผู้ตอบก็มีผลเป็นอย่างมากต่อการไม่ตอบแบบสอบถาม (Auriat, 1991) พบว่า ผู้สูงอายุและผู้มีการศึกษาน้อยจะมีอัตราไม่ตอบสูง (Deleeuw, 2001; Herzog and Rodger, 1992 and Huisman, 1999; อ้างอิงจาก Deleeuw, 2001; Huisman and Van der Zouwen, 1998) พบอีกว่าหญิงและผู้ที่มีสุขภาพไม่ดีจะมีอัตราการไม่ตอบสูง บุคลิกภาพของผู้ตอบบางประการก็อาจจะมีผลต่ออัตราการไม่ตอบ ซึ่งอาจสูงบ้างต่ำบ้าง เช่น ผู้มีความเชื่อมั่นในตัวเองสูง ผู้มีความเป็นมิตร ผู้ไม่เก็บตัวและผู้ไม่มีนิสัยหยาบกระด้างจะมีอัตราการไม่ตอบคำถามต่ำ เป็นต้น (Huisman and Van der Zouwen, 1998)

2.1.8 ขนาดตัวอย่าง

จากการศึกษางานวิจัยที่เกี่ยวข้องกับข้อมูลสูญหาย การกำหนดขนาดตัวอย่าง พบว่าขนาดตัวอย่างที่ใช้ในการศึกษามีขนาดตั้งแต่ 50 ถึง 1,000 หน่วย แต่ส่วนใหญ่จะใช้ขนาดตัวอย่างในการศึกษา 100 200 และ 500 หน่วย ซึ่งเป็นขนาดตัวอย่างที่พบโดยทั่วไปในการวิจัยเชิงสำรวจ (จำลอง วงษ์ประเสริฐ, 2554)

การกำหนดขนาดตัวอย่างในการวิจัยเชิงสำรวจ ขนาดตัวอย่างมีความสำคัญ จะต้องกำหนดให้มีความเหมาะสม ถ้าใช้ตัวอย่างขนาดเล็กเกินไปก็จะทำให้การสรุปผลในการวิจัยครั้งนั้นมีความน่าเชื่อถือลดน้อยลง แต่ถ้าตัวอย่างมีขนาดใหญ่มากจนเกินไปก็จะทำให้ผู้วิจัยเสียเวลาและค่าใช้จ่ายที่สูง ดังนั้นผู้วิจัยจึงกำหนดขนาดตัวอย่างที่ใช้ในการศึกษาครั้งนี้เป็น 100 200 และ 500

2.1.9 ร้อยละข้อมูลสูญหาย

จากการศึกษางานวิจัยที่เกี่ยวข้องกับข้อมูลสูญหายทั้งในและต่างประเทศ ใช้ร้อยละข้อมูลสูญหายตั้งแต่ร้อยละ 5 ถึงร้อยละ 30 แต่พบว่าส่วนใหญ่นิยมใช้ร้อยละข้อมูลสูญหายร้อยละ 5 10 15 และ 20 แต่ในการศึกษาครั้งนี้ผู้วิจัยได้กำหนดร้อยละข้อมูลสูญหาย คือ ร้อยละ 5 10 15 20 และ 30 (จริยา แสงสุวรรณ, 2551; Song and other, 2005)

2.2 การประมาณค่าข้อมูลสูญหายแบบ Hot-Deck

วิธี Hot-Deck หรือเรียกอีกอย่างหนึ่งว่าวิธี Neighborhood Next Door หรือ Nearest Neighbor หรือ Closest Fit หรือ Nearby Household (Saurabh, 2003) เป็นวิธีทดแทนข้อมูลที่สูญหายด้วยข้อมูลของผู้ให้ข้อมูล (Donor) จากงานวิจัย/สำรวจเรื่องเดียวกัน เรียกว่า Hot (ตรงข้ามกับการนำข้อมูลจากแฟ้มอื่นแต่เป็นงานสำรวจเรื่องเดียวกันในครั้งก่อน เรียกว่า Cold Deck) คำว่า Deck หมายถึง ปึกของแบบสอบถามที่เราคัดออกมาจากกองตามวิธี Nearest Neighbor Search Technique เพื่อให้ได้หน่วยสำรวจที่มีคุณสมบัติคล้ายคลึงกับหน่วยที่มีข้อมูลสูญหาย อาจโดยวิธีกำหนดคุณสมบัติ (Categorization) หรือโดยการหาระยะทาง (Euclidian Distance) วิธี Hot-Deck เป็นวิธีที่สำนักงานสำมะโนประชากรของสหรัฐ อเมริกา และของแคนาดาใช้สร้างข้อมูลทดแทนรายการที่สูญหาย (Dinardo and Lemienx, 1997; Brunel, 2000; Hirsch and Schumacher, 2001; Geoffrey and Ferrado, 1994) โดยปกติเราจะนำวิธี Hot-Deck ไปใช้สร้างข้อมูลทดแทนในกรณีข้อมูลสูญหายโดยสุ่ม (Missing at Random, MAR)

ซึ่งเป็นรูปแบบการสูญหายที่ถือว่าข้อมูลที่สูญหายผูกพันอยู่กับบุคลิก ลักษณะของผู้ให้ข้อมูล (Ford, 1983; Rivzi, 1983; อ้างอิงจาก Roth and Switzer, 1995; และ Switzer and others, 1998) เช่น ข้อมูลที่สูญหายผูกพันอยู่กับอายุของผู้ตอบแบบสอบถาม พบว่า ผู้สูงอายุยินดีตอบข้อถามมากกว่าผู้ที่มีอายุน้อย หรือข้อมูลที่สูญหายผูกพันอยู่กับลักษณะของการครอบครองที่อยู่อาศัย พบว่าผู้ที่มีบ้านเป็นของตนเองคือมีฐานะเป็นเจ้าของบ้านยินดีตอบแบบสอบถามมากกว่าผู้เช่า (Paullin and Ferraro, 1994)

วิธี Hot-Deck มีข้อดีหลายประการที่ไม่อาจมองข้าม ในการศึกษาทางการแพทย์สามารถสร้างข้อมูลด้านสุขภาพและด้านอื่น ๆ ของผู้ป่วย (Sven, 2001) และในการศึกษาความเกี่ยวเนื่องระหว่างการทำหมันชายกับการเกิดมะเร็งในต่อมลูกหมากโดยสอบถามความเห็นของแพทย์ว่าความวิตกกังวลเรื่องนี้ของประชาชนจะมีผลต่อจำนวนการมาขอรับบริการผ่าตัดทำหมันชายหรือไม่ การศึกษาครั้งนี้นักวิจัยใช้วิธี Hot-Deck สร้างข้อมูล ทดแทนให้แก่รายการที่สูญหาย (Magnani, 1999)

ในการสำรวจเกี่ยวการใช้จ่ายของผู้บริโภค (Consumer Expenditure Survey, CE) ปรากฏว่ามีครัวเรือนจำนวนมากจำรายการค่าใช้จ่ายต่าง ๆ ไม่ได้ หรือแกล้งบอกค่าใช้จ่ายไว้ต่ำกว่าความเป็นจริงซึ่งแตกต่างจากผู้อื่นอย่างผิดสังเกต หรือโครงการสำรวจระดับชาติอื่น ๆ ที่ต่างก็พบปัญหาเรื่องข้อมูลสูญหายเช่นกัน เช่น การสำรวจรายจ่ายของครัวเรือน การสำรวจการเปลี่ยนแปลงประชากร (Current Population Change Survey, CPS) การสำรวจครัวเรือนเกี่ยวกับปัญหายาเสพติด (National Household Survey on Drug Abuse, NHSDA) การสำรวจเกี่ยวกับค่าจ้าง การสำมะโนประชากรของประเทศสหรัฐอเมริกา ปี ค.ศ. 2000 โดยสำนักงานสำมะโนประชากรแห่งสหรัฐอเมริกา (US Bureau of Census) การสำรวจแรงงานโดยสำนักงานสถิติแรงงาน (Bureau of Labor Statistics) และการสำมะโนประชากรโดยสำนักงานสำมะโนประชากรแห่งสหราชอาณาจักร ต่างก็ใช้วิธี Hot-Deck สร้างข้อมูลทดแทน (Roth and Switzer, 1995; Paullin and Ferraro, 1994; Hirsch and S chumacher, 2001; Dinardo and Lemieux, 2001; Brunel, 2000)

คำว่า Hot-Deck หมายถึง การนำข้อมูลจากภายในแฟ้มเดียวกันมาใช้เป็นข้อมูลทดแทนให้แก่รายการข้อมูลที่มีค่าสูญหาย ตรงข้ามกับคำว่า Cold-Deck ซึ่งเป็นกรณีที่เรานำเอาข้อมูลจากแฟ้มอื่น หรือข้อมูลในอดีตที่เชื่อถือได้มาใช้เป็นข้อมูลทดแทนให้แก่รายการข้อมูลที่สูญหาย ดังนั้น Hot-Deck จึงเป็นเรื่องของการไปนำเอาข้อมูลจากหน่วยสำรวจอื่นที่มีคุณสมบัติคล้ายคลึงกับหน่วยที่มีข้อมูลสูญหาย เรียกว่าผู้ให้ข้อมูล (Donor) มาแทนที่ในรายการข้อมูลที่สูญหายของหน่วยรับข้อมูล เป็นวิธีแทนที่ตัวแปรที่มีค่าสูญหายด้วยข้อมูลของตัวแปรอื่นที่คล้ายคลึงกัน อาจใช้วิธีแทนที่รายการที่มีข้อมูลสูญหายของหน่วยรับข้อมูลด้วยข้อมูลของหน่วยให้ข้อมูลที่มีระยะทาง (เราใช้ระยะทางเป็นดัชนีวัดความสัมพันธ์) ห่างกันสั้นที่สุด เรียกว่า การคัดเลือกตามระดับความสัมพันธ์กับผู้ให้ข้อมูล หรือใช้ข้อมูลจากหน่วยอื่น ๆ ที่มีคุณสมบัติทางประชากรตรงกัน เช่น ใช้ข้อมูลจากคนในกลุ่มที่มี เพศ อายุ การศึกษา อาชีพ รายได้ เดียวกัน เป็นข้อมูลทดแทนให้แก่รายการที่สูญหาย เรียกว่า การคัดเลือกตามเงื่อนไขกลุ่มย่อย (Categorization)

วิธี Hot-Deck มีข้อดีคือ ใช้ง่าย เข้าใจง่าย ใช้ข้อมูลจริงเป็นค่าทดแทน และมีความหลากหลายในวิธีการ การนิยามคุณสมบัติของผู้ให้ข้อมูลเป็นอย่างอื่นก็จะได้ Hot-Deck วิธีใหม่ ด้วยเหตุนี้วิธี Hot-Deck จึงมีได้มากมายหลากหลายวิธีให้เลือกใช้ ข้อเสียก็คือวิธีนี้เป็นวิธีเชิงปฏิบัติ

จึงยังไม่มีทฤษฎีสนับสนุนมากนัก ตัวแบบต่าง ๆ ผ่านการพัฒนาโดยการทดลองโดยวิธีจำลองแบบ มิใช่โดยวิธีคณิตสถิติ แม้จะยืดหยุ่นและเปิดกว้าง แต่ก็อาจมีข้อทักท้วงเรื่องหลักทางทฤษฎี

อนึ่ง ในกรณีที่เราเลือกใช้วิธี Hot-Deck โดยวิธีคัดเลือกตามเงื่อนไขกลุ่ม พบว่ามีจุดทักท้วง 2 ประการคือ (มนตรี พิริยะกุล, 2548)

1. จำนวนกลุ่มย่อยอาจมีมากมายจนยากแก่การควบคุม และแตกต่างกันไปตามแนวทางที่นักวิจัยแต่ละคนจะกำหนดขึ้นตามที่ตนเห็นว่าเหมาะสม การเพิ่มจำนวนตัวแปรจำแนกขึ้นจะทำให้ได้จำนวนกลุ่มย่อยมากขึ้น ประชากรในกลุ่มมีความคล้ายคลึงกันมากขึ้น การลดตัวแปรจำแนกลงก็จะลดจำนวนกลุ่มย่อยลง ลดความคล้ายคลึงของประชากรในกลุ่มย่อยลง

2. เมื่อเราประสงค์จะใช้ตัวแปรอื่นที่มิใช่ตัวแปรกลุ่มตามธรรมชาติมาร่วมเป็นตัวแปรจำแนก เราจะต้องแปลงรหัสของตัวแปรนั้นให้เป็นรหัสกลุ่ม เช่น ประสงค์จะใช้อายุร่วมเป็นตัวแปรจำแนกและเราได้บันทึกข้อมูลอายุไว้เป็นตัวเลขเชิงปริมาณ กรณีนี้อาจจำเป็นต้องแปลงรหัสเดิมเป็นรหัสใหม่ ดังนี้

อายุต่ำกว่า 20 ปี ให้รหัสใหม่เป็นหมายเลข 1
อายุต่ำกว่า 20-29 ปี ให้รหัสใหม่เป็นหมายเลข 2
อายุต่ำกว่า 30-39 ปี ให้รหัสใหม่เป็นหมายเลข 3
อายุต่ำกว่า 40-49 ปี ให้รหัสใหม่เป็นหมายเลข 4
อายุต่ำกว่า 50 ปีขึ้นไป ให้รหัสใหม่เป็นหมายเลข 5

การแปลงรหัสของตัวแปรจากรหัสละเอียดมาเป็นรหัสที่หยาบกว่าย่อมเป็นการสูญเสียสารสนเทศและการประมาณค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลในตารางย่อยกระทำได้ยาก

วิธี HDD (Hot-Deck Deterministic) วิธี HDD ดำเนินการเป็น 2 ขั้นตอนดังนี้

ขั้นตอนที่ 1 คำนวณหาระยะทางระหว่างหน่วยสำรวจที่มีข้อมูลสูญหายกับหน่วยที่มีข้อมูลครบทุกตัวแปรตามแนวทางของ Pairwise Deletion

ระยะทางคือดัชนีหนึ่งที่ใช้วัดความสัมพันธ์ระหว่างวัตถุ (Case) นิยมใช้มากในงานวิเคราะห์กลุ่ม (Cluster Analysis) เหตุที่ต้องวัดความเกี่ยวข้องด้วยระยะทางแทนที่จะวัดด้วยสหสัมพันธ์ เพราะค่าสหสัมพันธ์เป็นดัชนีวัดความเกี่ยวข้องกันระหว่างตัวแปรไม่อาจใช้กับวัตถุหรือหน่วยสำรวจได้ วัตถุหรือหน่วยสำรวจคู่ใดมีระยะทางห่างกันน้อยแสดงว่าวัตถุหรือหน่วยสำรวจนั้นมีคุณสมบัติใกล้เคียงกันมาก ขณะเดียวกันวัตถุหรือหน่วยสำรวจคู่ใดมีระยะทางห่างกันมากแสดงว่าวัตถุคู่นั้นมีคุณสมบัติต่างกันมาก

ขั้นตอนที่ 2 นำข้อมูลจากหน่วยสำรวจที่สมบูรณ์ (คือตอบทุกคำถาม) ที่สัมพันธ์กับหน่วยที่มีข้อมูลสูญหายมากที่สุด (ระยะทางสั้นที่สุด) เป็นข้อมูลทดแทนแก่รายการที่สูญหายของหน่วยที่มีข้อมูลสูญหาย

ถ้าพบว่าหน่วยที่มีข้อมูลสูญหายสัมพันธ์กับหน่วยที่สมบูรณ์หลายหน่วย โดยวัดเป็นระยะทางที่เท่า ๆ กันได้หลายค่า ให้เลือกใช้ข้อมูลจากหน่วยที่อยู่ใกล้กว่า (มีเลขที่หน่วยสำรวจใกล้กันมากกว่า) เป็นข้อมูลทดแทน โดยมีความเชื่อว่าหน่วยที่มีเลขที่ใกล้เคียงกันคือหน่วยสำรวจในย่านหรือในพื้นที่ทางภูมิศาสตร์หรือกลุ่มประชากรย่อยเดียวกัน น่าจะมีธรรมชาติคล้ายกันกว่าหน่วยที่อยู่ต่างพื้นที่

ตาราง 2.1 ข้อมูลสมมติ

คนที่	เพศ	รายได้	อายุ	VAR1	VAR2	VAR3	VAR4	VAR5
1	1	1	3	4	2	4	5	4
2	1	2	2	.	3	4	3	4
3	2	3	4	3	.	.	4	5
4	1	3	3	.	4	3	2	4
5	2	1	2	4	2	2	.	.
6	1	3	2	.	3	5	5	.
7	1	2	4	5	4	3	3	4
8	2	3	3	5	4	3	.	3
9	1	1	1	5	4	4	.	5
10	2	2	5	5	4	5	4	5

จากแฟ้มข้อมูลเดิม พบว่า หน่วยสำรวจที่ 1 7 และ 10 เป็นหน่วยสำรวจสมบูรณ์ เราจะใช้หน่วยทั้งสามนี้เป็นหน่วยให้ข้อมูล (Donor) โดยทำหน้าที่ทดแทนข้อมูลให้แก่หน่วยที่ 2 3 4 5 6 8 9 แต่จะนำข้อมูลจากหน่วยใดไปเป็นข้อมูลทดแทนแก่หน่วยใดให้พิจารณาจากระยะทาง ซึ่งเป็นดัชนีที่ใช้วัดระดับความสัมพันธ์ระหว่างวัตถุ

โดยที่ระยะทางเราคำนวณโดยใช้สูตรทางคณิตศาสตร์ ดังนี้

$$d^2_{vv'} = \sum_{i \in obs(v)} (x_{vi} - x_{v'i})^2$$

v หมายถึง หน่วยสำรวจที่มีข้อมูลสูญหายไปบางตัวแปร

v' หมายถึง หน่วยสำรวจที่มีข้อมูลครบทุกตัวแปร

obs (v) หมายถึง จำนวนข้อมูลที่มีอยู่ทั้งหมด (เฉพาะคำตอบของคนที่ v)

$i \in$ obs (v) หมายถึง การคำนวณให้กระทำเฉพาะข้อมูลที่สอดคล้องกันของหน่วยสำรวจที่ v และ v'

ค่าระยะทางระหว่างหน่วยสำรวจคู่ใดที่สั้นกว่าแสดงว่าหน่วยสำรวจคู่นั้นมีความใกล้ชิดกันมากกว่า ค่าระยะทางที่น้อยลง แสดงว่าหน่วยสำรวจคู่นั้นยิ่งมีความละม้ายคล้ายคลึงกันมากขึ้น ค่าระยะทางระหว่างหน่วยสำรวจคู่ใดที่สูงขึ้นแสดงว่าหน่วยสำรวจคู่นั้นมีความใกล้ชิดกันน้อยลง จากข้อมูลในตาราง 2.1 ที่ยกมาข้างต้นเราสามารถคำนวณหาค่าของระยะทางได้ดังตาราง 2.2

ตาราง 2.2 ค่าระยะทางระหว่างหน่วยสำรวจตามวิธี Pairwise Deletion

	คนที่ 1	คนที่ 2	คนที่ 3	คนที่ 4	คนที่ 5	คนที่ 6	คนที่ 7	คนที่ 8	คนที่ 9	คนที่ 10
คนที่ 1	0.00	2.24	1.73	3.74	2.00	1.41	3.16	2.65	2.45	2.83
คนที่ 2	2.24	0.00	1.41	1.73	2.24	2.24	1.41	1.73	1.41	2.00
คนที่ 3	1.73	1.41	0.00	2.24	1.00	1.00	2.45	2.83	2.00	2.00
คนที่ 4	3.74	1.73	2.24	0.00	2.24	3.74	1.00	1.00	1.41	3.00
คนที่ 5	2.00	2.24	1.00	2.24	0.00	3.16	2.45	2.45	3.00	3.74
คนที่ 6	1.41	2.24	1.00	3.74	3.16	0.00	3.00	2.24	1.41	1.41
คนที่ 7	3.16	1.41	2.45	1.00	2.45	3.00	0.00	1.00	1.41	2.45
คนที่ 8	2.65	1.73	2.83	1.00	2.45	2.24	1.00	0.00	2.24	2.83
คนที่ 9	2.45	1.41	2.00	1.41	3.00	1.41	1.41	2.24	0.00	1.00
คนที่ 10	2.83	2.00	2.00	3.00	3.74	1.41	2.45	2.83	1.00	0.00

จากตารางระยะทาง เราจะตรวจสอบทีละคนโดยดูจากหน่วยสำรวจที่มีข้อมูลสูญหายคนแรก ในที่นี้คือคนที่ 2 ว่าสัมพันธ์กับคนที่มีข้อมูลครบทุกข้อคือคนที่ 1 7 และ 10 อย่างไร (แสดงในตารางด้วยเลขตัวทึบ) และพบว่า

คนที่ 2 สัมพันธ์กับคนที่ 7 มากกว่าคนที่ 1 และคนที่ 10 ให้นำข้อมูลตัวแปรตัวที่ 1 ของคนที่ 7 ไปเป็นข้อมูลทดแทนให้คนที่ 2

คนที่ 3 สัมพันธ์กับคนที่ 1 ที่สุด ให้นำข้อมูลของตัวแปรที่ 2 และตัวแปรที่ 3 ของคนที่ 1 ไปเป็นข้อมูลทดแทนให้แก่คนที่ 3

คนที่ 4 สัมพันธ์กับคนที่ 7 ที่สุด ให้นำข้อมูลของตัวแปรที่ 1 ของคนที่ 7 ไปเป็นข้อมูลทดแทนให้แก่คนที่ 4

คนที่ 5 คนที่ 6 คนที่ 8 และคนที่ 9 ก็ดำเนินการเช่นเดียวกัน ผลการทดแทนข้อมูลจะปรากฏแฟ้มข้อมูลดังตาราง 2.3 ตัวเลขที่ขีดเส้นใต้คือข้อมูลทดแทน

ตาราง 2.3 ข้อมูลและข้อมูลทดแทนตามวิธี Hot-Deck

คนที่	เพศ	รายได้	อายุ	VAR1	VAR2	VAR3	VAR4	VAR5
1	1	1	3	4	2	4	5	4
2	1	2	2	5	3	4	3	4
3	2	3	4	3	2	4	4	5
4	1	3	3	5	4	3	2	4
5	2	1	2	4	2	2	5	4
6	1	3	2	5	3	5	5	5
7	1	2	4	5	4	3	3	4
8	2	3	3	5	4	3	3	3
9	1	1	1	5	4	4	4	5
10	2	2	5	5	4	5	4	5

2.3 การประมาณค่าข้อมูลสูญหายแบบ Corrected Item Mean Substitution (CIM)

วิธี CIM (Corrected Item Mean Substitution) วิธี CIM เป็นวิธีที่นำเอาทั้งผลกระทบจากบุคคลและผลกระทบจากคำถามมาร่วมเป็นตัวถ่วงน้ำหนักเพื่อใช้สร้างข้อมูลทดแทน ผลกระทบจากบุคคลคือยอดรวมคำตอบของบุคคล ถ้าผู้ตอบรายใดตอบคำถามมากคำถามกว่า น้ำหนักของบุคคลผู้นั้นอาจสูงกว่าน้ำหนักของผู้อื่นที่ตอบคำถามน้อยคำถามกว่า ถ้าตัวแปรใดมีข้อมูลสูญหายน้อยกว่าตัวแปรนั้นย่อมมีน้ำหนักมากกว่าตัวแปรอื่นที่มีข้อมูลสูญหายมากกว่า

จากแฟ้มข้อมูล เราสามารถคำนวณหาค่าเฉลี่ยและยอดรวมทั้งของตัวแปรและของผู้ตอบได้ดังนี้ ค่าเฉลี่ยที่ปรากฏในตารางต่อไปนี้เป็นค่าเดียวกันกับค่าเฉลี่ยที่กล่าวแล้วในวิธี IMS และ PMS (มนตรี พิริยะกุล, 2548)

ตาราง 2.4 ข้อมูล ค่าเฉลี่ยของตัวแปร และยอดรวมตัวแปรตามของบุคคล (Person Sum)

คนที่	เพศ	รายได้	อายุ	VAR1	VAR2	VAR3	VAR4	VAR5	รวม
1	1	1	3	4	2	4	5	4	**19**
2	1	2	2	.	3	4	3	4	**14**
3	2	3	4	3	.	.	4	5	**12**
4	1	3	3	.	4	3	2	4	**13**
5	2	1	2	4	2	2	.	.	**8**
6	1	3	2	.	3	5	5	.	**17**

ตาราง 2.4 (ต่อ)

คนที่	เพศ	รายได้	อายุ	VAR1	VAR2	VAR3	VAR4	VAR5	รวม
7	1	2	4	5	4	3	3	4	**19**
8	2	3	3	5	4	3	.	3	**15**
9	1	1	1	5	4	4	.	5	**18**
10	2	2	5	5	4	5	4	5	**23**
เฉลี่ย				**4.43**	**3.33**	**3.67**	**3.71**	**4.25**	**19.39**

วิธีการคำนวณค่าทดแทนข้อมูลสูญหายของวิธี CIM สามารถคำนวณจากสมการ ดังต่อไปนี้

$$CIM_{vi} = w_v \bar{x}_{.i}^{(i)} = \left(\frac{\sum x_{vi}}{\sum_{i \varepsilon obs(v)} \bar{x}_{.i}^{(i)}} \right) \bar{x}_{.i}^{(i)} \quad ; v = 1, 2, \ldots, m_i$$

$$= \frac{\bar{x}_{.i}^{(i)} \times \sum_{i \varepsilon obs(v)} x_{vi}}{\sum_{i \varepsilon obs(v)} \bar{x}_{.i}^{(i)}} \quad ; v = 1, 2, \ldots, m_i$$

โดยที่ $\bar{x}_{.i}^{(i)}$ หมายถึง ค่าเฉลี่ยของตัวแปรที่ i คำนวณมาจากเฉพาะข้อมูลของคนที่ตอบข้อถามข้อที่ I สัญลักษณ์ (i) กำกับเอาไว้เพื่อแสดงว่าสนใจเฉพาะผู้ที่ตอบคำถามข้อที่ i

obs (v) หมายถึง บรรดาคำตอบของผู้ตอบรายที่ v สำหรับคำถามต่าง ๆ เฉพาะข้อที่ตอบ

$i \in$ obs (v) หมายถึง เราสนใจเฉพาะคำตอบของทุกคำถามที่ตอบ คำถามที่ไม่ตอบไม่นำมาพิจารณา

การทดแทนข้อมูลให้ทดแทนคราวละตัวแปร ดังนี้

ตัวแปรที่ 1 ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 2 ด้วย (14/19.39) x 4.43 = 3.20
ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 4 ด้วย (13/19.39) x 4.43 = 2.97
ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 6 ด้วย (17/19.39) x 4.43 = 3.88
ตัวแปรที่ 2 ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 3 ด้วย (12/19.39) x 3.33 = 2.06
ตัวแปรที่ 3 ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 3 ด้วย (12/19.39) x 3.67 = 2.27
ตัวแปรที่ 4 ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 6 ด้วย (8/19.39) x 3.71 = 1.53
ตัวแปรที่ 4 ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 8 ด้วย (15/19.39) x 3.71 = 2.87
ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 9 ด้วย (18/19.39) x 3.71 = 3.44

ตัวแปรที่ 5 ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 5 ด้วย (8/19.39) x 4.25 = 1.75
ทดแทนข้อมูลให้แก่ผู้ตอบคนที่ 6 ด้วย (17/19.39) x 4.25 = 3.73

เมื่อแทนที่ข้อมูลที่สูญหายในแฟ้มข้อมูลด้วยค่าทดแทนตามสูตรที่ปรากฏท้ายตารางจะปรากฏผลดังต่อไปนี้ ตัวเลขที่ขีดเส้นใต้คือข้อมูลทดแทนที่คำนวณมาโดยวิธีถ่วงน้ำหนัก

ตาราง 2.5 ข้อมูลและข้อมูลทดแทนตามวิธี CIM

คนที่	เพศ	รายได้	อายุ	VAR1	VAR2	VAR3	VAR4	VAR5
1	1	1	3	4	2	4	5	4
2	1	2	2	3.20	3	4	3	4
3	2	3	4	3	2.06	2.27	4	5
4	1	3	3	2.97	4	3	2	4
5	2	1	2	4	2	2	1.53	1.75
6	1	3	2	3.88	3	5	5	3.73
7	1	2	4	5	4	3	3	4
8	2	3	3	5	4	3	2.87	3
9	1	1	1	5	4	4	3.44	5
10	2	2	5	5	4	5	4	5

วิธี CIM เป็นวิธีถ่วงน้ำหนักให้กับข้อมูลที่จะใช้ในการประมาณค่าของรายการข้อมูลที่สูญหาย ค่าประมาณนี้นำเอาทั้งลักษณะจำเพาะของหน่วยสำรวจ ธรรมชาติของกลุ่มที่ตอบสนองเฉพาะข้อถาม และธรรมชาติของกลุ่มที่ตอบสนองแบบสอบถามทั้งชุด หมายความว่าการทดแทนข้อมูลที่สูญหายของประเด็นใดแก่หน่วยสำรวจใดนั้น ให้พิจารณาลักษณะเฉพาะตัวของผู้ตอบร่วมไปกับลักษณะการตอบสนองเฉพาะปัญหาหรือเฉพาะประเด็นของกลุ่ม คือ สนใจที่จะพิจารณาทั้งแนวความคิดเห็นส่วนตัวของผู้ตอบและแนวความคิดเห็นโดยรวมของกลุ่มเฉพาะประเด็น ไม่แยกสนใจเฉพาะแนวความคิดเห็นส่วนตัวของผู้ตอบหรือเฉพาะแนวความคิดเห็นโดยรวมของกลุ่ม การสนใจเฉพาะแนวความเห็นส่วนตัวของผู้ตอบคือวิธี PMS การสนใจเฉพาะประเด็นปัญหาตามแนวความคิดของกลุ่มต่อประเด็นนั้นคือวิธี IMS นอกจากนี้วิธี CIM ยังได้นำเอาความเห็นโดยรวมของกลุ่มต่อเรื่องที่สำรวจมาร่วมพิจารณาด้วย วิธี CIM จึงเป็นวิธีผสมระหว่างวิธี IMS และวิธี PMS

จากการศึกษาพบว่าวิธี CIM เป็นวิธีที่ดีที่สุด เนื่องจากให้ค่า Root Mean Square Deviation (RMSD) ต่ำที่สุด ที่ตัวอย่างขนาด 100 200 และ 400 ตามลำดับ สัดส่วนค่าสูญหายของข้อมูลเท่ากับ 0.05 0.12 และ 0.20 ตามลำดับ (Huisman, 1997)

2.4 งานวิจัยที่เกี่ยวข้อง

2.4.1 งานวิจัยในประเทศ

เชาว์ อินใย (2547: 305 – 314) ได้ศึกษาการพัฒนาวิธีการประมาณข้อมูลสูญหายแบบอีพีเอสอีและตรวจสอบความแม่นยำ และอำนาจการทดสอบที่ได้จากวิธีการประมาณข้อมูลสูญหายแบบอีพีเอสอีกับวิธีการประมาณข้อมูลสูญหายด้วยวิธีอีเอ็ม และการตัดข้อมูลสูญหายแบบลิสท์ไวส์ ตามวิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่ม และแบบหลายขั้นตอน ที่ระดับความสัมพันธ์ระหว่างตัวแปรระดับต่ำ (r = .30) ปานกลาง (r = .50) และสูง (r = .70) และจำนวนข้อมูลสูญหายร้อยละ 5 10 20 และ 30 และศึกษาปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีจัดการข้อมูลสูญหาย สัดส่วนข้อมูลสูญหาย และระดับความสัมพันธ์ระหว่างตัวแปรที่มีต่อความแม่นยำของการประมาณค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ ข้อมูลที่ใช้ศึกษามีลักษณะการแจกแจงแบบปกติสองตัวแปร และใช้วิธีมอนติคาร์โล ผลการศึกษาพบว่า วิธีการจัดการข้อมูลสูญหายโดยการแทนค่าแบบอีพีเอสอี ได้ค่าความแม่นยำของการประมาณค่าเฉลี่ยเลขคณิตไม่แตกต่างจากวิธีการประมาณข้อมูลสูญหายด้วยวิธีอีเอ็ม และมีความแม่นยำสูงที่สุดเมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่ม และแบบหลายขั้นตอน สัดส่วนข้อมูลสูญหายสูงสุดร้อยละ 30 ระดับความสัมพันธ์ระหว่างตัวแปรสูง การตัดข้อมูลสูญหายแบบลิสท์ไวส์ ได้ค่าความแม่นยำของการประมาณค่าความแปรปรวนแตกต่างจากวิธีอื่นอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และมีความแม่นำสูงที่สุดเมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่ม และแบบหลายขั้นตอน สัดส่วนข้อมูลสูญหายร้อยละ 10 20 และ 30 ทุกระดับความสัมพันธ์ระหว่างตัวแปร และในการประมาณค่าสัมประสิทธิสหสัมพันธ์ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ ได้ค่าความแม่นยำของการประมาณค่าสัมประสิทธิ์สหสัมพันธ์แตกต่างจากวิธีอื่นอย่างมีนัยสำคัยทางสถิติที่ระดับ .05 และมีความแม่นยำสูงที่สุดเมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่มและแบบหลายขั้นตอน สัดส่วนข้อมูลสูญหายสูงสุดร้อยงะ 30 ระดับความสัมพันธ์ระหว่างตัวแปรสูง วิธีการประมาณข้อมูลสูญหายแบบอีพีเอสอีได้ค่าอำนาจการทดสอบ เมื่อใช้การทดสอบความสัมพันธ์สูงกว่าการตัดข้อมูลสูญหายแบบลิสท์ไวส์และวิธีการประมาณข้อมูลสูญหาด้วยวิธีอีเอ็ม เมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น สัดส่วนข้อมูลสูญหายทุกระดับ ระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ไม่มีปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการประมาณข้อมูลสูญหาย สัดส่วนข้อมูลสูญหายและระดับความสัมพันธ์ระหว่างตัวแปรที่มีต่อความแม่นยำของการประมาณค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ แต่พบปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการประมาณข้อมูลสูญหาย และสัดส่วนข้อมูลสูญหาย ปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการประมาณข้อมูลสูญหาย และระดับความสัมพันธ์ระหว่างตัวแปร ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย สัดส่วนข้อมูลสูญหายและระดับความสัมพันธ์ระหว่างตัวแปรที่มีต่อความแม่นยำของการประมาณค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

มนตรี พิริยะกุล (2548: 117-122) ได้ทำการศึกษาเรื่อง ตัวแบบทดแทนข้อมูลในการวิจัยทางสังคมศาสตร์ : การจำลองแบบกรณีตัวแบบอย่างง่าย จำนวน 18 วิธีทั้งในกรณีเมื่อใช้กับข้อมูลจริงและเมื่อใช้กับข้อมูลเลียนแบบ โดยศึกษาเฉพาะในสถานการณ์ที่มีข้อมูลสูญหายไปโดยสุ่ม (Missing at Random, MAR) กำหนดให้มีสัดส่วนหน่วยสำรวจไม่สมบูรณ์ (m_c) รวม 4 ระดับคือ 5% 10% 15%

และ 20% กำหนดให้มีสัดส่วนข้อถามที่ไม่ได้รับการตอบสนอง (m_i) รวม 3 ระดับคือ 10% 20% และ 30% กำหนดให้มีขนาดตัวอย่าง (n) รวม 3 ขนาดคือ 100 200 และ 500 แต่ละเงื่อนไขการทดลองทำการทดลอง 100 ซ้ำ เฉพาะในกรณีของการทดลองมอนติคาร์โลจะเพิ่มเงื่อนไขเกี่ยวกับระดับของภาวะร่วมเส้นตรงพหุรวม 5 ระดับลงในเงื่อนไขการทดลองด้วย พบว่า วิธีทดแทนข้อมูลสูญหายที่ดีจำแนกได้เป็น 2 กลุ่มคือ กลุ่มที่ 1 กรณีเมื่อข้อมูลวัดค่าตามมาตรานามบัญญัติจะประกอบด้วยวิธี IMS วิธี HDD-IMS และวิธี HDD-IMS-Z ตามลำดับ และกลุ่มที่ 2 กรณีข้อมูลวัดค่าตามมาตราอันดับแบบลิกเกอร์ทจะประกอบด้วยวิธี PMS วิธี HDD-PMS และวิธี HDD-PMS-Z ตามลำดับ

จริยา แสงสุวรรณ และคณะ (2551: 1 – 7) ได้ศึกษาเปรียบเทียบวิธีประมาณค่าสูญหายในการวิเคราะห์การวิเคราะห์การถดถอยพหุคูณ โดยทำการเปรียบเทียบวิธีประมาณค่าสูญหาย 4 วิธี คือ วิธีสูญหาย วิธีค่าเฉลี่ย วิธีการสมการถดถอย และวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า โดยใช้ค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองเป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ โดยใช้เงื่อนไขต่าง ๆ ดังนี้ ขนาดตัวอย่าง 50 70 100 และ 200 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนร้อยละ 1 5 และ 15 สัดส่วนข้อมูลสูญหายร้อยละ 5 10 20 และ 30 ระดับความสัมพันธ์ระหว่างตัวแปร 3 ระดับ คือ ต่ำ (.20) ปานกลาง (.50) และสูง (.70) โดยทำการจำลองข้อมูลโดยใช้เทคนิคมอนติคาร์โล ทำซ้ำจำนวน 5,000 ครั้ง ผลการศึกษาพบว่า วิธีการสมการถดถอยและวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่ามีประสิทธิภาพใกล้เคียงกัน แต่วิธีการสมการถดถอยมีความง่ายและสะดวกในการใช้งานมากกว่า จึงเป็นวิธีที่เหมาะสมสำหรับการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอย

เพียงออ ยี่สา (2551: เว็บไซต์) ศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุเพื่อการพยากรณ์โดยพิจารณาข้อมูล 2 ลักษณะ คือ ข้อมูลภาคตัดขวาง และข้อมูลอนุกรมเวลาที่มีปัจจัยด้านแนวโน้ม และปัจจัยฤดูกาลเข้ามาเกี่ยวข้องโดยทำการประมาณค่าสูญหายด้วยวิธี Regression Imputation (RI) วิธี Nearest Neighbor Imputation (NNI) วิธี Weighted Nearest Neighbor and Regression Imputation (WNR) และวิธี EM Algorithm (EM) เกณฑ์ในการเปรียบเทียบประสิทธิภาพในการประมาณค่าสูญหายจะใช้ค่า MAPE ผลการวิจัยสรุปได้ดังนี้ สำหรับข้อมูลภาคตัดขวาง กรณีที่ค่าสหสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระทั้ง 2 ตัวสูง เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำถึงปานกลางวิธี RI และ EM มีค่า MAPE ต่ำกว่าวิธีอื่น เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับสูงวิธี WNR ให้ผลดีกว่าวิธีอื่น ๆ ที่นำมาเปรียบเทียบ กรณีที่ค่าสหสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระตัวหนึ่งสูงมากและอีก ตัวหนึ่งปานกลาง วิธี RI และ EM จะให้ผลดีกว่าวิธีอื่น ๆ สำหรับข้อมูลอนุกรมเวลา วิธี WNR มีค่า MAPE ต่ำกว่าวิธีการประมาณค่าอื่น ๆ ในกรณีที่ข้อมูลที่มีอิทธิพลของฤดูกาลสูง และวิธี NNI จะให้ผลดีเมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับสูง สำหรับข้อมูลที่มีอิทธิพลจากปัจจัยแนวโน้มสูงวิธี RI และ EM เป็นวิธีที่ให้ผลดีกว่าวิธีอื่น ๆ ที่นำมาเปรียบเทียบ กรณีที่ข้อมูลมีอิทธิพลจากปัจจัยแนวโน้มและปัจจัยฤดูกาลระดับปานกลาง เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำ วิธี RI และ EM มีค่า MAPE ต่ำกว่าวิธีการประมาณค่าอื่น ๆ เมื่อส่วนเบี่ยงเบนมาตรฐานเพิ่มสูงขึ้น วิธี WNR เป็นวิธีที่มีค่า MAP E ต่ำกว่าวิธีการประมาณค่าอื่น ๆ ที่นำมาเปรียบเทียบ

จิรกานต์ นวลละออง และเสาวณิต สุขภารังสี (2552: 106 – 119) ได้ศึกษาวิธีการประมาณข้อมูลสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุ เมื่อข้อมูลเป็นภาคตัดขวาง (Cross Sectional Data) และเมื่อข้อมูลเป็นอนุกรมเวลา (Time Series Data) เพื่อศึกษาวิธีประมาณค่าสูญหายของตัวแปรตามด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยการถดถอย วิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ วิธีการประมาณข้อมูลสูญหายร่วมด้วย ตัวถ่วงน้ำหนัก (Combined Missing Values Estimation Method) โดยจำลองข้อมูลด้วยเทคนิคมอนติคาร์โลด้วยโปรแกรม R ซ้ำ 50,000 ครั้ง และกำหนดขนาดตัวอย่าง 30 50 และ 100 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1 2 5 และ 10 ร้อยละของข้อมูลสูญหายเท่ากับ 5 10 15 และ 20 ขนาดความสัมพันธ์ของตัวแปรอิสระ คือ 0 0.2 และ 0.5 เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ พบว่า สำหรับทุกกรณีข้อมูลภาคตัดขวางในทุกกรณีนั้น เมื่อพิจารณาจากค่า MAPE วิธีการประมาณข้อมูลสูญหายเดี่ยวแล้ว วิธีเดี่ยวเลือกใช้วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยเป็นวิธีการที่เหมาะสมสุดในการประมาณข้อมูลสูญหาย สำหรับวิธีการประมาณข้อมูลสูญหายนั้น วิธีร่วมเลือกวิธีค่าสัมบูรณ์ต่ำสุดเป็นวิธีการที่เหมาะสมสุดในการประมาณค่าสูญหาย สำหรับข้อมูลอนุกรมเวลาเมื่อพิจารณาจากค่า MAPE สำหรับกรณีองค์ประกอบฤดูกาลสูงวิธีเดี่ยวเลือกวิธีกำลังสองน้อยสุด ยกเว้นในกรณีที่เมื่อขนาดตัวอย่างเท่ากับ 30 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 5 และขนาดตัวอย่างเท่ากับ 30 และ 50 สำหรับกรณีที่ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 10 นั้น เหมาะสมกับวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณค่าสูญหายร่วมเหมาะสมกับตัวถ่วงน้ำหนักร่วมวิธีค่าสัมบูรณ์ต่ำสุด เมื่อข้อมูลมีองค์ประกอบแนวโน้มสูงและข้อมูลที่มีองค์ประกอบแนวโน้มและฤดูกาลปานกลาง วิธีเดี่ยวเลือกใช้วิธีกำลังสองน้อยสุด และวิธีประมาณค่าสูญหายร่วมเหมาะสมกับตัวถ่วงน้ำหนักด้วยวิธีค่าสัมบูรณ์ต่ำสุด

จำลอง วงษ์ประเสริฐ (2554: บทคัดย่อ) ได้ศึกษาการพัฒนาวิธีการข้อมูลสูญหายโดยการถ่วงน้ำหนักแบบวนซ้ำด้วยวิธีของแจ็คไนฟ์และการวิเคราะห์การถดถอย (IWJR) และเปรียบเทียบประสิทธิภาพในการประมาณค่าเฉลี่ย ความแปรปรวน สัมประสิทธิ์สหสัมพันธ์และอำนาจการทดสอบ กับการตัดข้อมูลสูญหายแบบลิสท์ไวส์ (LD) วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (MI) วิธีการประมาณข้อมูลสูญหายด้วยการถดถอย (RI) และวิธีการจัดการข้อมูลสูญหายแบบอีพีอาร์ (EPR) โดยใช้ข้อมูลจากการจำลองและข้อมูลจริง ซึ่งมีขนาดตัวอย่าง 3 ขนาด (100 200 และ 500) ระดับความสัมพันธ์ระหว่างตัวแปร 3 ระดับ (ต่ำ ปานกลาง และสูง) ร้อยละของข้อมูลสูญหาย 4 ระดับ (ร้อยละ 5 10 15 และ 20) ผลวิจัยปรากฏดังนี้ ในการประมาณค่าเฉลี่ยเมื่อขนาดตัวอย่าง 100 และ 200 วิธีการ RI และ IWJR มีประสิทธิภาพสูงสุด เมื่อขนาดตัวอย่าง 500 วิธีการ LD EPR และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ วิธีการ MI RI และ IWJR มีประสิทธิภาพสูงที่สุด เมื่อระดับความสัมพันธ์ระหว่างตัวแปรปานกลางและสูง วิธีการ RJ และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าความแปรปรวนเมื่อระดับความสัมพันธ์ระหว่างตัวแปรต่ำ ปานกลางและสูง วิธีการ LD และ IWJR มีประสิทธิภาพสูงที่สุด ในการประมาณค่าเฉลี่ย ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ประชากร วิธีการ IWJR มีความแกร่งต่อขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรและร้อยละข้อมูลสูญหาย ซึ่งสอดคล้องกันทั้งข้อมูลการจำลองและข้อมูลจริง

บวรวรรณ ดิเรกโภค (2543: เว็บไซต์) ได้ศึกษาการประยุกต์ใช้วิธีใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าในโรงพยาบาลราชวิถี เมื่อตัวแปรมีข้อมูลสูญหาย โดยใช้จำนวนค่าที่ใส่แทนค่าสูญหายแบบเชิงสุ่ม M=3, 5 กับการใส่ค่าเพียงค่าเดียว M=1 โดยใช้ข้อตกลงเบื้องต้นว่าข้อมูลมีการสูญหาย แบบเชิงสุ่ม (MAR) ตัวอย่างการศึกษาคือข้อมูลทุติยภูมิผู้ขับขี่จักรยานยนต์ที่บาดเจ็บ เข้ารับการรักษาที่แผนกฉุกเฉิน โรงพยาบาลราชวิถี จำนวน 2,668 ราย ตั้งแต่วันที่ 1 มกราคม – 31 ธันวาคม 2538 พบว่า ค่าความคลาดเคลื่อนของสัมประสิทธิ์การถดถอยแบบลอจิสติกส์เมื่อใช้ M=3, 5 มีค่ามากกว่าการใส่ค่าแทนข้อมูลที่สูญหายเพียงค่าเดียว M=1 ในโมเดลการถดถอยแบบลอจิสติกส์ ซึ่งมีตัวแปรตามคือ ระดับคะแนนโคม่าปกติ ส่วนตัวแปรอิสระได้แก่ การใช้ยา ที่มีผลต่อการควบคุมพาหนะขณะขับขี่ การใช้หมวกนิรภัย การดื่มเครื่องดื่มผสมแอลกอฮอล์ของผู้ขับขี่ อายุ และเพศ ซึ่งตัวแปรสามตัวแรกเป็นตัวแปรร่วมทวิที่มีค่าสูญหาย เมื่อเปรียบเทียบค่าประมาณในระหว่างชุดข้อมูลที่ใส่ค่าหลายค่าพบค่าความ คลาดเคลื่อนของสัมประสิทธิ์การถดถอยแบบลอจิสติกส์ของตัวแปรอิสระเมื่อใช้ M=3 มีค่ามากกว่า M=5 ยกเว้นตัวแปรการใช้หมวกนิรภัย และความคลาดเคลื่อนของสัมประสิทธิ์ การถดถอยแบบลอจิสติกส์จากการวิเคราะห์ชุดข้อมูลที่สมบูรณ์มีค่าสูงกว่าข้อมูลที่ใส่ค่าแทนข้อมูลที่สูญหายแต่ละค่า จากการวิเคราะห์ชุดข้อมูลที่ใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าแสดงให้เห็นว่าโมเดลการถดถอยแบบลอจิสติกส์มีนัยสำคัญทางสถิติ วิธีนี้ดีกว่าการวิเคราะห์ข้อมูลที่มีเพียงหน่วยตัวอย่างซึ่งได้ค่าสังเกตของตัวแปรครบ ทุกตัวเท่านั้น และวิธีนี้สามารถประยุกต์ใช้ได้กับข้อมูลที่ตัวแปรไม่ได้รับคำตอบ ซึ่งจะได้ผลดีกับตัวแปรชนิดแบ่งกลุ่ม

2.4.2 งานวิจัยต่างประเทศ

Crawford และคณะ (1995: 15 – 20) ได้ทำการเปรียบเทียบวิธีการประมาณข้อมูลสูญหาย 4 วิธี คือ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ วิธีการประมาณข้อมูลสูญหายด้วยค่าเดียว วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ โดยใช้ข้อมูลโครงการสุขภาพคนชรา โดยทำการเปรียบเทียบภายใต้ประเภทของข้อมูลสูญหายต่างกัน ผลการศึกษาพบว่า การตัดข้อมูลสูญหายแบบลิสท์ไวส์ และวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย ค่าประมาณที่ได้มีความเอนเอียง วิธีประมาณข้อมูลสูญหายด้วยค่าเดียวและวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย จะให้ค่าความแปรปรวนต่ำกว่าความเป็นจริง ส่วนวิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ ค่าประมาณค่าเฉลี่ยประชากรเป็นค่าประมาณที่ไม่เอนเอียง

Roth และคณะ (1996: 101 – 112) ได้ศึกษาโดยทำการเปรียบเทียบวิธีการประมาณข้อมูลสูญหาย 4 วิธี คือ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ การตัดข้อมูลสูญหายแบบแพร์ไวส์ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการประมาณข้อมูลสูญหายด้วยการถดถอยในบริบทของการวิเคราะห์การถดถอย การศึกษาครั้งนี้ใช้ข้อมูลเกี่ยวกับการจัดการทรัพยากรมนุษย์ โดยใช้ขนาดตัวอย่างเท่ากับ 3,169 ผลจากการศึกษาพบว่า เมื่อทำการประมาณข้อมูลสูญหายด้วยวิธีการทั้ง 4 วิธี การทดสอบระดับความสัมพันธ์ระหว่างตัวแปรที่ใช้ในการศึกษาไม่สอดคล้องกัน เช่น การทดสอบมีนัยสำคัญในกรณีที่ทำการประมาณข้อมูลสูญหายด้วยวิธีการตัดข้อมูลสูญหายแบบแพร์ไวส์ แต่ไม่มีนัยสำคัญในวิธีการตัดข้อมูลสูญหายแบบลิสท์ไวส์ แต่การศึกษาในครั้งนี้ไม่ได้ทำการสรุปว่าวิธีการใดมีประสิทธิภาพมากที่สุด

Guertin (1968: 896) ได้วิจัยเชิงประจักษ์และการศึกษาของ Couper (1966) ทำการเปรียบเทียบวิธีการประมาณข้อมูลสูญหาย 3 วิธี คือ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยและวิธีการประมาณข้อมูลสูญหายด้วยการถดถอย โดยใช้ขนาดตัวอย่าง 463 ร้อยละข้อมูลต่าง ๆ ที่สูญหายในแต่ละตัวแปรตั้งแต่ร้อยละ 2 ถึงร้อยละ 55 จากการศึกษาของ Guertin พบว่า วิธีการทั้งสามวิธีให้ผลที่ได้ในการวิเคราะห์สหสัมพันธ์แตกต่างกันทั้งสามวิธี

Jackson (1968: 835 – 844) ได้ทำการเปรียบเทียบวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยและวิธีการประมาณข้อมูลสูญหายด้วยการถดถอย โดยใช้จำนวนตัวอย่าง 7,084 ตัวอย่าง ข้อมูลสูญหายร้อยละ 10 15 และ 20 พบว่า วิธีการประมาณข้อมูลสูญหายด้วยการถดถอยมีประสิทธิภาพมากกว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย

Huisman (2000: 331-351) ได้ศึกษาวิธีแทนที่ข้อมูลสูญหายโดยใช้ข้อมูลจาก 4 ชุดข้อมูล คือ FFPI (E) RAND (PF) RAND (MH) และ NHPR (SI) ตัวอย่างมีขนาด 100 200 และ 400 ตามลำดับ สัดส่วนค่าสูญหายของข้อมูลเท่ากับ 0.05 0.12 และ 0.20 ตามลำดับ โดยนำวิธี Incorrect Answer Substitution(IAS), Random Draw Substitution (RDS), Item Mean Substitution (IMS), Person Mean Substitution (PMS), Corrected Item Mean Substitution (CIM), Item Correlation Substitution (ICS), Hot-Deck Random (HDR), Hot-Deck Deterministic (HDD) และวิธี Hot-Deck Next Case (HNC) นำมาทดสอบเปรียบเทียบกัน พบว่า วิธี CIM เป็นวิธีที่ดีที่สุดเนื่องจากให้ค่า Root Mean Square Deviation (RMSD) ต่ำที่สุด

Viragoontavan (2000: 116 – 121) ได้ศึกษาเปรียบเทียบประสิทธิภาพสัมพัทธ์ (Relative Effectiveness) ของวิธีจัดการข้อมูลสูญหาย 6 วิธี คือ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ วิธีประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยการถดถอย วิธีการประมาณข้อมูลสูญหายแบบฮอทเดสก์ วิธีการประมาณข้อมูลสูญด้วยวิธีเอ็มไอ ด้วยโปรแกรม SOLAS และโปรแกรม NORM ภายใต้บริบทของการวิเคราะห์จำแนกประเภท โดยใช้การจำลองข้อมูลที่มีเงื่อนไขต่าง ๆ ดังนี้ ระดับความสัมพันธ์ระหว่างตัวแปร ต่ำ ปานกลาง และสูง ขนาดตัวอย่าง 100 200 และ 500 และสัดส่วนข้อมูลสูญหาย .05 .10 และ .20 รูปแบบของการสูญหายคือ การสูญหายแบบสุ่ม จากการศึกษาพบว่า การตัดข้อมูลสูญหายแบบลิสท์ไวส์มีประสิทธิภาพต่ำสุด วิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ ด้วยโปรแกรม SOLAS และโปรแกรม NORM มีประสิทธิภาพสูงที่สุด วิธีการประมาณข้อมูลสูญหายทั้ง 6 วิธี เมื่อข้อมูลมีความสัมพันธ์กันน้อย ประมาณค่าได้ถูกต้องมากกว่าเมื่อข้อมูลมีความสัมพันธ์กันสูง

Strike และคณะ (2001: 890-908) ได้ทดลองโดยใช้วิธีแทนที่ข้อมูลสูญหายด้วยวิธี Listwise Deletion, Mean Imputation และวิธี Hot Deck อีก 8 วิธี รวมทั้งสิ้น 10 วิธี โดยใช้ข้อมูลจาก National Health and Nutrition Examination Survey (NHANES) จำนวน 800 ตัวอย่าง และวิธีที่ดีที่สุดจะต้องมีค่าความเอนเอียงต่ำและมีความแม่นยำสูง จากการทดลองพบว่า วิธี Hot Deck ที่ใช้การหาระยะทางแบบยูคลิด (Euclidean Distance) และคะแนนมาตรฐาน (Z-score Standardization) เป็นวิธีที่ให้ค่าความเอนเอียงต่ำสุดและมีความแม่นยำสูงสุด

Viviano and Lovision (2004: 4) ได้ศึกษาการประมาณข้อมูลสูญหายด้วยวิธีการอย่างง่าย และวิธีการตัวแปรพหุ ในการประมาณข้อมูลสูญหายในบริบทของการวิเคราะห์การถดถอยโลจีสติกส์ โดยใช้การจำลองสถานการณ์ด้วยมอนติคาร์โลเทคนิค พบว่า วิธีการอย่างง่ายและวิธีการตัวแปรพหุให้ประสิทธิภาพของการประมาณค่าในการถดถอยโลจีสติกส์ไม่แตกต่างกัน และพบว่า ถ้าข้อมูลสูญหายมีจำนวนมาก ค่าจริงของความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 มีค่าลดลง และค่าดังกล่าวขึ้นอยู่กับจำนวนข้อมูลสูญหาย การทดแทนข้อมูลสูญหายก่อนการทำการวิเคราะห์ข้อมูลทำให้ได้สารสนเทศเพิ่มขึ้นและได้ผลสรุปที่ถูกต้องมากขึ้น

Chaimongkol (2004: 104 – 107) ได้ศึกษาวิธีการประมาณข้อมูลสูญหาด้วยค่าถดถอดและเข้าใกล้สุดแบบถ่วงน้ำหนัก (Weighted Nearest Neighbor and Regression Imputation (WNR)) ในการประมาณค่าเฉลี่ยของประชากร โดยทำการเปรียบเทียบกับวิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์ไนจ์บอร์ดและวิธีการประมาณข้อมูลสูญหายด้วยการถดถอย พบว่า สัดส่วนข้อมูลสูญหายน้อยกว่าร้อยละ 15 และในทุกระดับความสัมพันธ์ระหว่างตัวแปร วิธีประมาณค่าข้อมูลสูญหายด้วยค่าถดถอยและเข้าใกล้สุดแบบถ่วงน้ำหนักมีค่าความเอนเอียงของการประมาณค่าเข้าใกล้ 0 ที่สัดส่วนข้อมูลสูญหายน้อยกว่าร้อยละ 15 ด้วยความเชื่อมั่นร้อยละ 95 สำหรับทุกขนาดตัวอย่างและสัดส่วนข้อมูลสูญหาย

Chaimongkol และ Suwattee (2004: 185 – 192) ได้พัฒนาวิธีประมาณค่าสูญหายด้วยการรวมแนวคิดของสองวิธีเข้าด้วยกัน ซึ่งเราเรียกว่า การประมาณข้อมูลสูญหายด้วยการถดถอยของค่าใกล้สุด (Nearest Neighbor Regression Imputation) จากนั้นทำการเปรียบเทียบกับวิธีอื่นที่มีอยู่แล้ว 3 วิธี คือ วิธีการประมาณข้อมูลสูญหายด้วยการถดถอย วิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์ไนจ์บอร์ และวิธีประมาณโดยใช้ค่าใกล้สุดของค่าถดถอย (Regression-Based Nearest Neighbor Hot-Decking) ผลการเปรียบเทียบไม่สามารถสรุปได้ว่าวิธีใดดีที่สุด เนื่องจากวิธีหนึ่งอาจจะเหมาะสมสำหรับข้อมูลหรือสิ่งที่ต้องการพิจารณาอย่างใดอย่างหนึ่งเท่านั้น อย่างไรก็ตาม วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าใกล้สุด เป็นวิธีที่เหมาะสมเมื่อพิจารณาจากเกณฑ์การวัดบางอย่างที่ใช้ในการศึกษาครั้งนี้

Hassani และคณะ (2004: 271 – 278) ได้ศึกษาเปรียบเทียบวิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์ไนจ์บอร์และวิธีการประมาณข้อมูลสูญหายด้วยการถดถอย โดยใช้สัดส่วนข้อมูลสูญหายร้อยละ 20 เปรียบเทียบโดยพิจารณาจากความเอนเอียง ค่าเบี่ยงเบนเฉลี่ยสัมบูรณ์ และรากที่สองของค่าเฉลี่ยความคลาดเคลื่อน พบว่า วิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์ไนจ์บอร์มีประสิทธิภาพมากที่สุด

Acock (2005: 1012 – 1028) ได้ศึกษาเปรียบเทียบวิธีประมาณค่าข้อมูลสูญหาย ได้แก่ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ การตัดข้อมูลสูญหายแบบแพร์ไวส์ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยและวิธีการประมาณข้อมูลสูญหายหลายวิธี (Multiple Imputation) พบว่าวิธีประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยเป็นวิธีการที่มีประสิทธิภาพต่ำที่สุด เนื่องจากให้ค่าประมาณค่าความแปรปรวนแย่ที่สุด การตัดข้อมูลสูญหายแบบลิสท์ไวส์ การตัดข้อมูลสูญหายแบบแพร์ไวส์มีประสิทธิภาพเป็นที่ยอมรับได้ ถ้าข้อมูลสูญหายมีจำนวนน้อย การสูญหายของข้อมูลเป็นการสูญหายแบบสุ่มสมบูรณ์ แต่อำนาจการทดสอบทางสถิติลดต่ำลง การประมาณค่าข้อมูลสูญหายด้วยค่าเดียว

ไม่เหมาะสมอย่างยิ่งในการประมาณค่าข้อมูลสูญหาย ควรใช้วิธีประมาณค่าข้อมูลสูญหายที่ได้จากการพิจารณาค่าที่ไม่ใช่ค่าเดียว

Velicer และ Colby (2005: 596 – 615) ได้ศึกษาการประมาณข้อมูลสูญหายกับการเก็บรวบรวมข้อมูลระยะยาวของข้อมูลอนุกรมเวลา โดยทำการเปรียบเทียบวิธีการประมาณข้อมูลสูญหาย 4 วิธี ได้แก่ การตัดข้อมูลสูญหายแบบลิสท์ไวส์ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยค่าที่อยู่ติดกันและการประมาณข้อมูลสูญหายด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) โดยสร้างสถานการณ์จำลองสร้างข้อมูลอนุกรมเวลาขนาด 100 คาบเวลา จำนวน 50 ชุดที่แตกต่างกัน โดยกำหนดเงื่อนไขระดับของสหสัมพันธ์ของตัวรบกวน (Autocorrelation) 5 ระดับ ความชัน 2 ระดับ และสัดส่วนของข้อมูลสูญหาย 5 ระดับ โดยทำการเปรียบเทียบการประมาณค่าพารามิเตอร์ 4 ตัว ได้แก่ ระดับความแปรปรวนของความคลาดเคลื่อน ระดับของสหสัมพันธ์ของตัวรบกวนและความชัน พบว่า การประมาณข้อมูลสูญหายด้วยวิธี ภาวะน่าจะเป็นสูงสุดให้ค่าที่ถูกต้องมากที่สุด 4 พารามิเตอร์ ทุก ๆ เงื่อนไขของการทดลอง โดยที่วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยให้ค่าที่ถูกต้องต่ำที่สุด

Huang และ Carriere (2006: 235 – 247) ได้พัฒนาวิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ โดยใช้วิธีการวัดซ้ำเปรียบเทียบวิธีการประมาณข้อมูลสูญหายในข้อมูลแบบวัดซ้ำในตัวอย่างขนาดเล็ก ภายใต้เงื่อนไขข้อตกลงเบื้องต้นของการแจกแจงปกติหลายตัวแปร โดยใช้ข้อมูลที่ได้จากการจำลอง และเปรียบเทียบวิธีประมาณสูญหายระหว่างวิธีประมาณสูญหายโดยใช้วิธีเม็กซิมั่มไลลิฮูด กับวิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ เพื่อทดสอบสมมติฐานอิทธิพลของทรีทเมนต์ และอิทธิพลของความคลาดเคลื่อนของทรีทเมนต์ ผลการศึกษาพบว่า วิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอกับวิธีประมาณข้อมูลที่สูญหายโดยวิธีเม็กซิมั่มไลลิฮูดจะให้ผลการทดสอบที่ดีกว่า

Qinbao และ Martin (2007: 51) ได้พัฒนาวิธีประมาณค่าข้อมูลสูญหายกับข้อมูลที่มีตัวอย่างจำนวนน้อย โดยได้นำวิธีคลาสมีนอิมพิวเตชัน (Class Mean Imputation : CMI) และเคเนียร์เรสท์ ไนจ์บอร์ฮ็อทเดสก์ (k-NN) มาผสานวิธีเรียกว่า เอ็มไอเอ็นไอ (MINI) โดยทำการทดลองกับขนาดตัวอย่าง 50 และ 100 โดยมีสัดส่วนของข้อมูลสูญหายร้อยละ 10 15 20 และ 30 โดยประเภทของข้อมูลสูญหายที่ใช้ในการทดลองคือ การสูญหายแบบสุ่มอย่างสมบูรณ์และการสูญหายแบบสุ่ม พบว่า วิธีเอ็มไอเนไอมีประสิทธิภาพสูงกว่าวิธีคลาสมีนอิมพิวเตชันและวิธีเคเนียร์รสท์ไนจ์บอร์ ฮ็อทเดสก์

Granberg-Rademacker (2007: 325 – 338) ได้ศึกษาเปรียบเทียบวิธีประมาณค่าข้อมูลสูญหายระหว่างการตัดข้อมูลสูญหายแบบลิสท์ไวส์ มาร์คอฟเชนมอนติคาร์โลกับขั้นตอนวิธีการเก็บตัวอย่างแบบกิ๊บ (Markov Chain Monte Carlo with the Gibbs Sampler Algorithm : MCMC) และการประมาณค่าข้อมูลสูญหายหลายวิธีโดยสมการลูกโซ่ (Multiple Imputation by Chained Equations : MICE) ภายใต้การสูญหายแบบสุ่มสมบูรณ์ การสูญหายแบบสุ่มและการสูญหายที่ไม่สามารถละเลยได้ (Missingness Nonignorable : IN) และข้อมูลสูญหายร้อยละ 5 10 และร้อยละ 15 โดยใช้เทคนิคมอนติคาร์โล และเปรียบเทียบประสิทธิภาพโดยใช้รากที่สองของค่าเฉลี่ยกำลังสอง (Root Mean Square Error) พบว่า วิธีการที่มี

ประสิทธิภาพสูงสุดในการประมาณค่าข้อมูลสูญหายคือ การประมาณค่าข้อมูลสูญหายหลายวิธีโดยสมการลูกโซ่ทุกเงื่อนไข มาร์คอพเชนมอนติคาร์โลกับขั้นตอนวิธีการเก็บตัวอย่างแบบกิ๊บมีเสถียรภาพในการประมาณค่าข้อมูลสูญหายสูงสุด ส่วนการตัดข้อมูลสูญหายออกแบบลิสท์ไวส์มีประสิทธิภาพต่ำสุดทุก ๆ เงื่อนไข

Walker (2008: 466 – 479) ได้ตรวจสอบผลของข้อมูลสูญหายในทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า ภายใต้เงื่อนไขการสูญหายแบบสุ่มอย่างสมบูรณ์และเปรียบเทียบวิธีการจัดการกับข้อมูลสูญหาย 4 วิธี ได้แก่ การตัดข้อมูลสูญหายแบบแพร์ไวส์ วิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์ไนจ์บอร์ วิธีการประมาณข้อมูลสูญหายแบบฮ็อทเดสก์ และวิธีการประมาณข้อมูลสูญหายโดยใช้แบบจำลอง ภายใต้แบบจำลองทฤษฎีการตอบสนองข้อสอบที่มี 2 พารามิเตอร์ พบว่า เมื่อสัดส่วนของข้อมูลสูญหายเพิ่มขึ้น การพยายามที่จะทำให้แบบจำลองสอดคล้องกับทฤษฎีเป็นไปได้ยากเพิ่มขึ้น และการวินิจฉัยรายบุคคลผิดพลาดเพิ่มขึ้น สำหรับวิธีจัดการข้อมูลสูญหายที่สามารถทำให้สามารถวินิจฉัยรายบุคคลได้ถูกต้องมากที่สุด คือ การตัดข้อมูลสูญหายแบบแพร์ไวส์ สำหรับผลกระทบของข้อมูลสูญหายที่เด่นชัดที่สุด คือ การแสดงสาเหตุที่ผิดพลาดในการวิเคราะห์เชิงสาเหตุและยังผลให้แบบจำลองจากการวัดไม่สอดคล้องกับแบบจำลองตามทฤษฎี

Andridge และ Little (2010: 40 - 64) ได้เปรียบเทียบประสิทธิภาพวิธีการประมาณข้อมูลสูญหายด้วยวิธีฮ็อทเดค วิธีนาอีฟและวิธีแจ็คไนฟ์ โดยใช้ข้อมูลจาก The Third National Health and Nutrition Examination Survey (NHANES III) จำนวน 800 ตัวอย่าง พบว่าวิธีการประมาณข้อมูลสูญหายแบบ HDD มีประสิทธิภาพสูงเมื่อใช้กับข้อมูลจริงที่ตัวอย่างมีขนาดใหญ่และร้อยละข้อมูลสูญหายมีจำนวนมาก

Ting (2010: 277 – 287) ได้เปรียบเทียบประสิทธิภาพวิธีการประมาณข้อมูลสูญหายด้วยวิธีอีเอ็มและมาร์คอฟ เชน มอนติแคร์โล (MCMC) โดยทำการเปรียบเทียบความถูกต้องของการทดแทนข้อมูลสูญหาย โดยใช้ข้อมูลเชิงประจักษ์และข้อมูลที่ได้จากการจำลอง จากกาดรศึกษาไม่พบความแตกต่างของประสิทธิภาพในการประมาณข้อมูลสูญหายของสองวิธี และสัดส่วนข้อมูลสูญหายและอัตราการสูญหายของตัวแปรไม่มีอิทธิพลต่อประสิทธิภาพของวิธีการประมาณข้อมูลสูญหายด้วยวิธีอีเอ็มและวิธีการประมาณข้อมูลสูญหายด้วยเทคนิค มาร์คอฟ เชน มอนติคาร์โล

Gabriel และคณะ (2010: 1 – 10) ได้ทำการเปรียบเทียบประสิทธิภาพวิธีการประมาณข้อมูลสูญหายกับข้อมูลทางด้านจิตวิทยา วิธีการที่ทำการเปรียบเทียบคือ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ และการประมาณค่าแบบเอฟไอเอ็มแอล จากการศึกษาพบว่า วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยเป็นวิธีที่มีประสิทธิภาพต่ำที่สุด และถ้าต้องการประมาณค่าข้อมูลสูญหายวิธีการที่แนะนำคือ วิธีการประมาณข้อมูลสูญหายด้วยวิธีเอ็มไอ และการประมาณค่าแบบเอฟไอ เอ็มแอล โดยการเลือกใช้ให้พิจารณาที่รูปแบบการสูญหายของข้อมูล

James (2010: 223 – 243) ได้ศึกษาอัตราการตอบสนองของประสิทธิภาพของการศึกษาระยะยาวในการสำรวจด้วยตัวอย่าง โดยทำการศึกษาจากข้อมูลเชิงประจักษ์เพื่อหาความเอนเอียงที่เกิดขึ้นจากการไม่ตอบสนองต่อข้อคำถามในแบบสอบถาม และได้พัฒนาวิธีการวัดแบบใหม่เพื่อลดความเสี่ยงจากการไม่ตอบสนองต่อข้อคำถามเรียกว่า เศษส่วนของสารสนเทศที่สูญหาย (The Fraction of Missing Information : FMI) เพื่อจัดการกับข้อมูลสูญหาย โดยพยายาม

ลดความไม่แน่นอนของอัตราการไม่ตอบสนองต่อข้อคำถาม โดยวิธีการที่พัฒนาขึ้นใช้เป็นเครื่องมือในการตรวจสอบในขั้นตอนการเก็บรวบรวมข้อมูลเพื่อให้สารสนเทศจากข้อมูลที่เก็บรวบรวมสูงสุด

จากการศึกษางานวิจัยในประเทศและต่างประเทศ พบว่า งานวิจัยเกี่ยวกับการจัดการข้อมูลสูญหายส่วนใหญ่จะศึกษาโดยใช้เทคนิคมอนติคาร์โลจำลองข้อมูลที่ใช้ในการศึกษา และปัจจัยที่ส่งผลต่อประสิทธิภาพการประมาณข้อมูลสูญหาย ได้แก่ ระดับความสัมพันธ์ระหว่างตัวแปร ขนาดตัวอย่าง ร้อยละข้อมูลสูญหาย และวิธีจัดการข้อมูลสูญหาย มักจะทดแทนข้อมูลสูญหายด้วยค่าเพียงค่าเดียวที่ประมาณได้ วิธีการที่ใช้จะมี 2 ลักษณะ คือ ใช้ข้อมูลที่ไม่สูญหายในตัวแปรนั้นในการประมาณค่าเพื่อทดแทนข้อมูลสูญหาย เช่น วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย และใช้วิธีการพิจารณาตัวแปรอื่นที่มีอยู่ที่มีความสัมพันธ์กับตัวแปรที่มีข้อมูลสูญหาย ในการพิจารณาค่าทดแทนข้อมูลสูญหาย เช่น วิธีการประมาณข้อมูลสูญหายแบบฮ็อทเดคก์ วิธีการประมาณข้อมูลสูญหายแบบเนียร์เรสท์ไนจ์บอร์ฮ็อทเดคก์ วิธีการประมาณข้อมูลสูญหายด้วยการถดถอย และวิธีการดังกล่าวไม่ได้นำสารสนเทศที่ได้จากการประมาณค่าข้อมูลสูญหายมาพิจารณาในการประมาณค่าการทดแทนข้อมูลสูญหาย ซึ่งมีข้อดีและข้อบกพร่อง แสดงดังตาราง 2.6

ตาราง 2.6 ข้อดีและข้อบกพร่องของวิธีจัดการข้อมูลสูญหาย

วิธีการ	รายละเอียด	กรณีที่นำไปใช้	ข้อดี	ข้อบกพร่อง
การตัดข้อมูลสูญหายทิ้งไป - การตัดข้อมูลแบบลิสท์ไวส์ (Listwise Deletion)	ตัดค่าสังเกตหรือแถวที่มีข้อมูลสูญหายออกและนำค่าสังเกตที่สมบูรณ์ไปวิเคราะห์	ควรหลีกเลี่ยงในการนำไปใช้	วิธีการง่ายต่อการนำไปใช้	อำนาจการทดสอบลดลง
- การตัดข้อมูลแบบแพร์ไวส์ (Pairwise Deletion)	วิเคราะห์ข้อมูลจากข้อมูลส่วนที่มีค่าสมบูรณ์ทั้งสองตัวแปร โดยใช้ความสัมพันธ์ระหว่างตัวแปรคู่	เมื่อข้อมูลมีการสูญหายแบบสุ่มและข้อมูลสูญหายน้อยกว่า 10% ความสัมพันธ์ระหว่างตัวแปรมีความน่าเชื่อถือสูง	มีการใช้งานข้อมูลได้อย่างเต็มที่และมีประสิทธิภาพ สูญเสียอำนาจทดสอบน้อยกว่าการตัดข้อมูลแบบลิสท์ไวส์	อาจทำให้ค่าสหสัมพันธ์หรือค่าความแปรปรวนร่วมมีความเอนเอียง
การแทนที่ข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ - การประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean Imputation)	แทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย	เมื่อความสัมพันธ์ระหว่างตัวแปรต่ำและข้อมูลสูญหายน้อยกว่า 10%	มีการใช้งานข้อมูลได้อย่างเต็มที่และมีประสิทธิภาพและง่ายต่อการใช้งาน	ผลกระทบในทางลบต่อการประมาณค่าความแปรปรวนและระดับขององศาอิสระ
- การประมาณข้อมูลสูญหายด้วยวิธีฮอทเดค (Hot Deck Imputation)	แทนที่ข้อมูลสูญหายด้วยค่าคะแนนที่เกิดขึ้นจริงจากกรณีที่คล้ายคลึงกันในชุดข้อมูล	เมื่อข้อมูลสูญหายมีรูปแบบที่ชัดเจน	ข้อมูลที่สูญหายไปถูกแทนที่ด้วยค่าจริงและค่าเฉลี่ยของการกระจายไม่บิดเบือน	ทฤษฎีสนับสนุนยังน้อยและใช้ข้อมูลจริงเพื่อตรวจสอบความถูกต้อง, มีปัญหาหากไม่มีกรณีที่เกี่ยวข้องอื่น ๆ ในทุกด้านของชุดข้อมูล

ตาราง 2.6 (ต่อ)

วิธีการ	รายละเอียด	กรณีที่นำไปใช้	ข้อดี	ข้อบกพร่อง
- การประมาณข้อมูลสูญหายด้วยค่าถดถอย (Regression Imputation)	ประมาณค่าความสัมพันธ์ระหว่างตัวแปรและใช้ค่าสัมประสิทธิ์ประมาณค่าข้อมูลสูญหาย	เมื่อข้อมูลสูญหายมากกว่า 20% และตัวแปรมีความสัมพันธ์กันสูง	ประมาณค่าเบี่ยงเบนข้อมูลจากค่าเฉลี่ยและรูปร่างของการกระจาย	บิดเบือนค่าองศาอิสระและเพิ่มค่าความสัมพันธ์เทียม
การประมาณค่าพารามิเตอร์				
- การประมาณค่าควรจะเป็นสูงสุด (Maximum Likelihood)	ประมาณค่าพารามิเตอร์โดยข้อมูลที่มีอยู่และค่าข้อมูลถูกใช้ในการประมาณค่าข้อมูลสูญหายอีกครั้ง	เมื่อพบสมมติฐานการกระจายของข้อมูล	ความแม่นยำเพิ่มขึ้นถ้ารูปแบบถูกต้อง	สมมติฐานการกระจายของข้อมูลจำเป็นต้องใช้เทคนิคที่ค่อนข้างจะเข้มงวด
- การประมาณค่าคาดหวังสูงสุด (Expected Maximization)	ทำการวนซ้ำจนกว่าจะเกิดค่าที่ลู่เข้าหรือค่าที่เปลี่ยนแปลงน้อยมาก	เมื่อพบสมมติฐานการกระจายของข้อมูล	ความแม่นยำเพิ่มขึ้นถ้ารูปแบบถูกต้อง	อัลกอริทึมใช้เวลาในการมาบรรจบกันและมีความซับซ้อนเกินไป

ดังนั้นผู้วิจัยจึงได้นำเสนอวิธีจัดการข้อมูลสูญหายด้วยวิธี HDD-CIM กับวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี CIM และวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD การศึกษาโดยวิธีทดลองที่กระทำกับข้อมูลที่มีอยู่จริง โดยใช้ข้อมูลจากแฟ้มข้อมูลทุติยภูมิจากแฟ้มข้อมูลสำรวจความคิดเห็นของประชาชนเกี่ยวกับสถานการณ์การแพร่ระบาดยาเสพติด พ.ศ. 2554 ที่สำรวจโดยสำนักงานสถิติจังหวัดมหาสารคาม โดยแบ่งขั้นตอนการทำวิจัยเป็น 3 ขั้นตอน คือ 1) การพัฒนาวิธีประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM 2) การตรวจสอบประสิทธิภาพของวิธีการจัดการข้อมูลสูญหาย และ 3) การวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการจัดการข้อมูลสูญหาย ขนาดตัวอย่างและร้อยละของข้อมูลสูญหาย

บทที่ 1

บทนำ

1.1 ภูมิหลัง

ในการศึกษาวิจัยทั่ว ๆ ไปบ่อยครั้งที่ผู้วิจัยพบว่า ข้อมูลที่รวบรวมมาได้มีคำตอบไม่ครบถ้วนสมบูรณ์ ข้อมูลได้รับคำตอบเพียงบางส่วน ซึ่งส่งผลให้ไม่สามารถใช้ประโยชน์จากข้อมูลนั้นได้เต็มที่ ซึ่งปัญหาข้อมูลสูญหายอาจไม่ถือเป็นปัญหาที่รุนแรงหรืออาจถือว่าเป็นเรื่องเล็กน้อย ถ้าการวิเคราะห์ข้อมูลกระทำด้วยสถิติที่วิเคราะห์ข้อมูลรายตัวแปร เช่น ร้อยละ หรือสถิติพรรณนาอื่น แต่ถ้าหากการวิเคราะห์นั้นเป็นการวิเคราะห์ทางสถิติด้วยข้อมูลหลายตัวแปรจำเป็นต้องใช้วิธีวิเคราะห์เชิงพหุ (Multiple Analysis) การสูญหายของข้อมูลจะมีผลกระทบรุนแรง เพราะหน่วยวิเคราะห์ใดมีตัวแปรใดขาดข้อมูลไปแม้ตัวเดียวก็จะตัดหน่วยวิเคราะห์นั้นทิ้งทั้งหน่วย โดยไม่สนใจว่ายังมีตัวแปรอื่นที่มีข้อมูลครบถ้วนหรือไม่ (Heeringa, 2000: 1-19)

นักวิจัยจำเป็นต้องพิจารณาแนวทางที่เหมาะสม สำหรับใช้จัดการข้อมูลสูญหายในทุก ๆ ครั้งที่พบปัญหานี้ ซึ่งวิธีการที่ใช้สำหรับจัดการกับข้อมูลสูญหายมีทางเลือกให้พิจารณาค่อนข้างหลากหลาย หากเลือกใช้วิธีการจัดการกับข้อมูลสูญหายที่ไม่เหมาะสมย่อมส่งผลทำให้เกิดการบิดเบือนการวิเคราะห์ วิธีการทางสถิติได้ถูกพัฒนาเพื่อการวิเคราะห์ข้อมูลที่สมบูรณ์ แต่เมื่อมีข้อมูลสูญหายย่อมส่งผลกระทบต่อประสิทธิภาพการวิเคราะห์ข้อมูล จากการศึกษาของ Wood และคณะ (2004: 368-376) ที่ได้ทำการศึกษาจากผลงานวิจัยที่ได้รับการตีพิมพ์ในวารสารจำนวน 71 ชิ้น พบว่ามีงานวิจัยถึงร้อยละ 89 ที่มีปัญหาเรื่องข้อมูลสูญหาย และมีเพียงร้อยละ 21 เท่านั้นที่มีการจัดการกับข้อมูลสูญหายที่ไม่สมบูรณ์ นั่นแสดงให้เห็นว่า การจัดการกับปัญหาข้อมูลสูญหายยังคงถูกละเลยกันอย่างเป็นปกติ

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า นักวิจัยหลาย ๆ คน พยายามจัดการกับข้อมูลสูญหาย ซึ่งวิธีที่ใช้แก้ข้อมูลสูญหาย (Method for Treating Missing Data) สามารถแบ่งออกเป็น 3 วิธีการใหญ่ ๆ ได้ดังนี้

1. การตัดข้อมูลสูญหายทิ้งไป (Ignoring and Discarding Data) วิธีการนี้แบ่งออกเป็นสองวิธีดังนี้

1.1 การตัดข้อมูลแบบลิสท์ไวส์ (Listwise Data Deletion) เป็นวิธีการตัดค่าสังเกตหรือแถวที่มีข้อมูลสูญหายออกไปและนำค่าสังเกตที่สมบูรณ์เท่านั้นไปวิเคราะห์

1.2 การตัดข้อมูลแบบแพร์ไวส์ (Pairwise Data Deletion) เป็นวิธีที่ไม่ตัดแถวที่ขาดความสมบูรณ์ทิ้ง แต่จะนำแถวเหล่านั้นมาใช้ในการประมวลผลด้วย โดยพิจารณาทุก ๆ แถวที่มีค่าข้อมูลในตัวแปรที่กำลังสนใจ ถึงแม้ว่าตัวแปรอื่นจะมาสมบูรณ์ก็ตาม

2. การประมาณค่าพารามิเตอร์ (Parameter Estimation)

ตัวอย่างวิธีการประมาณค่า ได้แก่ การประมาณค่าความควรจะเป็นสูงสุด (Maximum Likelihood Stimation) ซึ่งคำนวณจากข้อมูลที่ทราบทั้งหมด ค่าความควรจะเป็นสูงสุดจะถูกคำนวณโดยแยกข้อมูลที่สมบูรณ์ของบางตัวแปรและคำนวณจากข้อมูลในทุกตัวแปร

ค่าความควรจะเป็นสูงสุดจากการคำนวณทั้งสอง นั่นคือ ข้อมูลที่สมบูรณ์ของบางตัวแปรและคำนวณจากข้อมูลในทุกตัวแปร จะถูกใช้ในการประมาณค่าข้อมูลสูญหายอีกทีหนึ่ง

3. การแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation)

วิธีการนี้เป็นวิธีที่ทำการประมาณค่าข้อมูลสูญหาย โดยอาศัยความสัมพันธ์ระหว่ากลุ่มข้อมูล เพื่อนำไปประมาณค่าข้อมูลสูญหาย ซึ่งวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณนั้น มีอยู่หลายวิธีด้วยกัน เช่น วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean Imputation หรือ MI) ข้อดีของวิธีการนี้ คือ มีการใช้งานข้อมูลได้อย่างเต็มที่และมีประสิทธิภาพ วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยโดยการถ่วงน้ำหนัก (Corrected Item Mean หรือ CIM) มีข้อดี คือ นำเอาทั้งผลกระทบจากบุคคลและผลกระทบจากคำถามมาร่วมเป็นตัวถ่วงน้ำหนักเพื่อใช้สร้างข้อมูลทดแทน วิธีการประมาณข้อมูลสูญหายด้วยค่าฮอทเดค (Hot Deck Deterministic Imputation หรือ HDD) มีข้อดี คือ ข้อมูลที่สูญหายไปถูกแทนที่ด้วยค่าจริงและค่าเฉลี่ยของการกระจายไม่บิดเบือน วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย (Regression Imputation หรือ RI) มีข้อดี คือ ประมาณค่าเบี่ยงเบนข้อมูลจากค่าเฉลี่ยและรูปร่างของการกระจาย และวิธีประมาณข้อมูลสูญหายด้วยค่าใกล้ที่สุด (Nearest Neighbor Imputation หรือ NNI) เป็นต้น

ดังนั้นในการศึกษาวิจัยครั้งนี้ ผู้วิจัยจึงนำวิธีการวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ยโดยการถ่วงน้ำหนัก (Corrected Item Mean หรือ CIM) และวิธีการประมาณข้อมูลสูญหายด้วยค่าฮอทเดค (Hot Deck Deterministic Imputation หรือ HDD) มารวมกันเพื่อสร้างวิธีการในการประมาณค่าข้อมูลสูญหายแบบใหม่ ที่มีความเอนเอียงต่ำและแม่นยำ และไม่เป็นภาระแก่นักวิจัยที่ไม่สันทัดวิธีการทางสถิติ

1.2 ความมุ่งหมายของการวิจัย

1.2.1 เพื่อสร้างวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่เทียบกับวิธี HDD และวิธี CIM

1.3 ความสำคัญของการวิจัย

1.3.1 เพื่อให้นักวิจัยได้มีแนวทางในการเลือกใช้วิธีการประมาณค่าข้อมูลสูญหาย ซึ่งปัจจุบันนักวิจัยยังขาดการเสนอแนะวิธีการที่จะช่วยในการเลือกวิธีการประมาณค่าข้อมูลสูญหายอย่างเพียงพอ ผู้วิจัยจึงได้นำเสนอวิธีการประมาณค่าข้อมูลสูญหายในการสำรวจอีกรูปแบบ โดยอาศัยข้อมูลที่สูญหายนั้นในการประมาณค่า เรียกวิธีนี้ว่า HDD-CIM ในการศึกษาครั้งนี้ ได้เจาะจงไปที่ปัจจัยต่าง ๆ ที่มีอิทธิพลต่อประสิทธิภาพของการจัดการข้อมูลสูญหาย เช่น ขนาดตัวอย่าง ร้อยละของข้อมูลสูญหาย รูปแบบของการสูญหายของข้อมูลที่ใช้ในการศึกษาครั้งนี้คือ การสูญหายแบบสุ่มสมบูรณ์ และเปรียบเทียบวิธีการประมาณค่าที่ได้จากการศึกษาครั้งนี้ กับวิธีการ CIM และวิธีการ HDD

1.3.2 เพื่อให้นักวิจัยได้มีวิธีการที่ง่ายในการจัดการข้อมูลสูญหาย ในกรณีที่ข้อมูลสูญหายนั้นเป็นข้อมูลสูญหายที่มีมาตราวัดแบบนามบัญญัติ (Nominal Scale) และมาตราวัดแบบเรียงอันดับ (Ordinal Scale)

1.4 ขอบเขตของการวิจัย

ขอบเขตของการวิจัยในครั้งนี้มีดังต่อไปนี้

1.4.1 ข้อมูลจริงที่ใช้เป็นข้อมูลทุติยภูมิ (Secondary Data) ที่ได้จากการสำรวจความคิดเห็นของประชาชนเกี่ยวกับสถานการณ์การแพร่ระบาดยาเสพติด พ.ศ. 2554 จาก 13 อำเภอ ที่สำรวจโดยสำนักงานสถิติจังหวัดมหาสารคาม จำนวน 2,975 ระเบียน เนื่องจากมีบางระเบียนที่ไม่สมบูรณ์ทำให้ได้จำนวนข้อมูลที่นำมาใช้ในการศึกษาครั้งนี้จำนวน 2,970 ระเบียน ซึ่งมีตัวแปร 8 ตัวแปร ดังนี้

a1 : เพศ
- 1 = ชาย
- 2 = หญิง

a2 : อายุ
- 1 = 18 – 19 ปี
- 2 = 20 – 29 ปี
- 3 = 30 – 39 ปี
- 4 = 40 – 49 ปี
- 5 = 50 – 59 ปี
- 6 = 60 ปีขึ้นไป

a3 : ระดับการศึกษาสูงสุด
- 1 = ไม่มีการศึกษา
- 2 = ต่ำกว่าศึกษา
- 3 = ประถมศึกษา
- 4 = มัธยมศึกษาตอนต้น
- 5 = มัธยมศึกษาตอนปลาย
- 6 = อุดมศึกษา
- 7 = อื่น ๆ

a4 : สถานภาพการทำงาน
- 1 = นายจ้าง
- 2 = ลูกจ้างรัฐบาล
- 3 = ลูกจ้างเอกชน
- 4 = ทำงานส่วนตัว
- 5 = ทำงานให้ครอบครัวโดยไม่ได้รับค่าจ้าง
- 6 = การรวมกลุ่ม

7 = อื่น ๆ

a5 : ปัจจุบันสถานการณ์การแพร่ระบาดยาเสพติด ในชุมชน/หมู่บ้านของท่าน เป็นอย่างไร

1 = ยังมีปัญหา

2 = ไม่มีปัญหา

3 = ไม่ทราบ/ไม่แน่ใจ

a6 : ปัจจุบันการหาซื้อยาเสพติดในชุมชน/หมู่บ้านของท่าน เป็นอย่างไร

1 = หาซื้อยาก

2 = หาซื้อง่าย

3 = หาซื้อไม่ได้

4 = ไม่ทราบ/ไม่แน่ใจ

a7 : ในชุมชน/ในหมู่บ้านของท่านมีเจ้าหน้าที่รัฐติดตามดูแล และให้ความสนใจในการป้องกันและแก้ไขปัญหายาเสพติดในชุมชน/หมู่บ้านของท่าน มากน้อยเพียงใด

1 = มากที่สุด

2 = มาก

3 = ปานกลาง

4 = น้อย

5 = ไม่ทราบ/ไม่แน่ใจ

a8 : ท่านพึงพอใจต่อผลการดำเนินงานของแกนนำ/ชมรมเกี่ยวกับการป้องกันและแก้ไขปัญหายาเสพติดในชุมชน/หมู่บ้านของท่านมากน้อยเพียงใด

0 = ไม่พอใจเลย

1 – 4 = น้อย

5 – 6 = ปานกลาง

7 – 8 = มาก

9 – 10 = มากที่สุด

1.4.2 การเปรียบเทียบประสิทธิภาพของวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM วิธีการประมาณข้อมูลสูญหายแบบ CIM และวิธีการประมาณข้อมูลสูญหายแบบ HDD โดยพิจารณาจากค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ค่าความเอนเอียง (Bias) (Roth and Stwitzer, 1995; Huisman, 1997) ค่าความแปรปรวน (Variance) และค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error : MAPE) ดังนี้

1.4.2.1 MSE = $\frac{1}{n}\sum(residual)^2$ โดยที่ residual = ค่าจริง – ค่าทดแทน

วิธีทดแทนข้อมูลใดให้ค่าเฉลี่ยของ MSE ต่ำกว่าแสดงว่าเป็นวิธีทดแทนข้อมูลที่มีความแม่นยำสูงกว่า

1.4.2.2 Bias = $\sqrt{\frac{1}{n}\sum(residual)}$ โดยที่ residual = ค่าจริง- ค่าทดแทน

วิธีทดแทนข้อมูลใดให้ค่าเฉลี่ยของ Bias ต่ำกว่าแสดงว่าเป็นวิธีทดแทนข้อมูลที่เอนเอียงน้อยกว่า

1.4.2.3 Variance = $\frac{\sum(x_i - \bar{x})^2}{n-1}$ โดยที่ x_i แทนค่าของข้อมูลที่ i , $\bar{x}$ แทนค่าเฉลี่ยข้อมูล และ n แทนจำนวนหน่วยตัวอย่าง

วิธีทดแทนข้อมูลใดให้ค่าเฉลี่ยของ variance ต่ำกว่าแสดงว่าเป็นวิธีทดแทนข้อมูลที่แปรปรวนน้อยกว่า

1.4.2.4 MAPE = $\frac{100}{n}\sum_{t=1}^{n}\left|\frac{x_t - \hat{x}_t}{x_t}\right|$ โดยที่ x_t แทนค่าจริง , $\hat{x}_t$ แทนค่าประมาณ , n แทนจำนวนข้อมูลสูญหาย

วิธีทดแทนข้อมูลใดให้ค่าเฉลี่ยของ MAPE ต่ำกว่าแสดงว่าเป็นวิธีทดแทนข้อมูลที่มีประสิทธิภาพสูงกว่า

1.4.3 ตัวแปรที่ศึกษามีดังต่อไปนี้

1.4.3.1 ตัวแปรอิสระ

1.4.3.1.1 วิธีจัดการข้อมูลสูญหาย จำแนกเป็น 3 วิธี คือ

1) วิธีการประมาณข้อมูลสูญหายแบบ CIM
2) วิธีการประมาณข้อมูลสูญหายแบบ HDD
3) วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM

1.4.3.1.2 ขนาดตัวอย่าง

1) ขนาดตัวอย่าง 100
2) ขนาดตัวอย่าง 200
3) ขนาดตัวอย่าง 500

1.4.3.1.3 ร้อยละข้อมูลสูญหาย

1) ร้อยละ 5
2) ร้อยละ 10
3) ร้อยละ 15
4) ร้อยละ 20
5) ร้อยละ 30

1.4.3.2 ตัวแปรตาม

1.4.3.2.1 ประสิทธิภาพของวิธีการประมาณข้อมูลสูญหาย

1.4.3.2.2 ปฏิสัมพันธ์ระหว่างวิธีการประมาณข้อมูลสูญหาย ขนาดตัวอย่าง และร้อยละข้อมูลสูญหาย

1.5. นิยามศัพท์เฉพาะ

1.5.1 ข้อมูลสูญหาย (Missing Data) หมายถึง ค่าสังเกตที่ต้องการทราบแต่ไม่สามารถทราบค่าได้ โดยที่ค่านั้นควรจะสามารถทราบค่าได้ หากวิธีการที่ใช้ในการรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีขึ้นหรือมีความเหมาะสมมากขึ้น

1.5.2 ความเอนเอียง (Bias) หมายถึง สถานการณ์ที่ค่าประมาณมิได้ชี้ไปที่ค่าจริงของสิ่งที่ต้องการศึกษา เป็นการมองในลักษณะเฉลี่ย ค่าประมาณอาจชี้ไปที่ใดก็ได้ ต่ำกว่าค่าจริงก็ได้สูงกว่าค่าจริงก็ได้

1.5.3 ค่าจริง หมายถึง ค่าที่ได้จากข้อมูลที่สมบูรณ์

1.5.4 ค่าประมาณ หมายถึง ค่าที่ได้จากวิธีการจัดการข้อมูลสูญหาย ด้วยวิธีการประมาณค่าข้อมูลสูญหาย

1.5.5 ประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหาย หมายถึง ความใกล้เคียงของค่าประมาณจากวิธีการประมาณค่าข้อมูลสูญหายกับค่าจริงจากข้อมูลสมบูรณ์ โดยใช้เกณฑ์ตัดสินประสิทธิภาพ คือ ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองและค่าความเอนเอียง

1.5.6 วิธีการประมาณข้อมูลข้อมูลสูญหาย หมายถึง การประมาณค่าของข้อมูลที่สูญหายไปเพื่อให้ข้อมูลที่เก็บรวบรวมมามีความสมบูรณ์ก่อนนำไปวิเคราะห์ ในการวิจัยครั้งนี้กำหนดวิธีการประมาณข้อมูลสูญหาย 3 วิธี คือ วิธีการประมาณข้อมูลสูญหายแบบ CIM วิธีการประมาณข้อมูลสูญหายแบบ HDD และวิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM

1.5.7 วิธีการประมาณข้อมูลสูญหายแบบ CIM หมายถึง วิธีที่นำเอาทั้งผลกระทบจากบุคคล (ผู้ตอบ) และผลกระทบจากคำถาม (ตัวแปร) มาร่วมเป็นตัวถ่วงน้ำหนักเพื่อใช้สร้างข้อมูลทดแทน

1.5.8 วิธีการประมาณข้อมูลสูญหายแบบ HDD หมายถึง วิธีที่รับข้อมูลสัมพันธ์กับหน่วยให้ข้อมูลหลายหน่วยคือมีระยะทางที่เท่ากันหลายค่าให้เลือกใช้ข้อมูลจากหน่วยให้ข้อมูลที่อยู่ใกล้กว่า (เลขที่หน่วยสำรวจอยู่ใกล้กันมากกว่า) เป็นข้อมูลทดแทน

1.5.9 วิธีการประมาณข้อมูลสูญหายแบบ HDD-CIM หมายถึง วิธีการประมาณข้อมูลสูญหายโดยขั้นตอนแรกทำการประมาณข้อมูลสูญหายโดยใช้วิธีการ CIM ขั้นตอนที่สองทำการประมาณค่าจากข้อมูลที่ได้จากขั้นตอนแรก โดยใช้วิธีการ HDD และขั้นตอนที่สามนำค่าที่ได้จากขั้นตอนที่สองมาแทนค่าในค่าข้อมูลที่สูญหาย

ประวัติย่อผู้วิจัย

ประวัติย่อผู้วิจัย

ชื่อนามสกุล	นายไพฑูรย์ มุลิวัลย์
วันเดือนปีเกิด	วันที่ 3 เมษายน พ.ศ. 2521
จังหวัด และประเทศที่เกิด	จังหวัดร้อยเอ็ด ประเทศไทย
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 60 หมู่ 12 ตำบลดงลาน อำเภอเมือง จังหวัดร้อยเอ็ด 45000
ประวัติการศึกษา	พ.ศ. 2539 มัธยมศึกษาตอนปลาย โรงเรียนพลาญชัยพิทยาคม จังหวัดร้อยเอ็ด พ.ศ. 2543 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาสถิติ มหาวิทยาลัยมหาสารคาม พ.ศ. 2556 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาวิทยาการจัดการสถิติ มหาวิทยาลัยมหาสารคาม
ตำแหน่ง สถานที่ทำงาน	นักวิทยาศาสตร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม ตำบลขามเรียง อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150
ที่อยู่ที่สามารถติดต่อได้	บ้านเลขที่ 60 หมู่ 12 ตำบลดงลาน อำเภอเมือง จังหวัดร้อยเอ็ด 45000

เอกสารอ้างอิง

เอกสารอ้างอิง

จริยา แสงสุวรรณ ประสิทธิ์ พยัคฆพงษ์ และอำไพ ทองธีรภาพ. (2551). “การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ,” *วารสารวิทยาศาสตร์ประยุกต์*, 7(1), 1 – 7; มิถุนายน.

จำลอง วงษ์ประเสริฐ. (2554). *การพัฒนาวิธีการประมาณข้อมูลสูญหายโดยการถ่วงน้ำหนักแบบวนซ้ำด้วยวิธีของแจ็คไนฟ์และการวิเคราะห์การถดถอย (ไอดับบลิวเจอาร์)*. วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต มหาสารคาม: มหาวิทยาลัยมหาสารคาม.

จิรกานต์ นวลละออง และเสาวณิต สุขภารังสี. (2553). “การเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับตัวแบบการพยากรณ์,” *เอกสารประกอบการประชุมทางวิชาการ สถิติและสถิติประยุกต์ ครั้งที่ 11 ประจำปี 2553*. ณ โรงแรมฮอลิเดย์อินน์ จังหวัดเชียงใหม่,

เชาว์ อินใย. (2547). *การพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีและการตรวจสอบความแม่นยำและอำนาจการทดสอบเปรียบเทียบกับวิธีอีเอ็มและลิสท์ไวท์ : เทคนิคมอนติคาร์โล*. วิทยานิพนธ์ปรัชญาการศึกษาดุษฎีบัณฑิต พิษณุโลก: มหาวิทยาลัยนเรศวร.

ณรงค์ โพธิ์ และสมชาย ปราการเจริญ. (2553). “การศึกษาการเปรียบเทียบความแม่นยำวิธีการประมาณค่าสูญหาย ระหว่างการวิเคราะห์จำแนกประเภทกับการวิเคราะห์ความถดถอยเชิงพหุคูณร่วมกับการวิเคราะห์องค์ประกอบ,” *การประชุมทางวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษา ครั้งที่ 11*. ขอนแก่น: มหาวิทยาลัยขอนแก่น.

บวรวรรณ ดิเรกโภค. (2543). *การประยุกต์วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าในการวิเคราะห์ ข้อมูล อุบัติเหตุผู้ขับขี่จักรยานยนต์*. วิทยานิพนธ์ปรัชญาวิทยาศาสตร์มหาบัณฑิต กรุงเทพฯ: มหาวิทยาลัยมหิดล.

ปิยะภรณ์ ประสิทธิ์วัฒนาเสรี และสุคนธ์ ประสิทธิ์วัฒนาเสรี. (2552). “ข้อมูลสูญหายและแนวทางการจัดการ,” *วารสารการจัดการข้อมูลและชีวสถิติ,* 4(3), 52-61.

เพียงออ ยีสา. (2551). *การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น*. วิทยานิพนธ์ปรัชญาวิทยาศาสตร์มหาบัณฑิต กรุงเทพฯ: จุฬาลงกรณมหาวิทยาลัย.

มนตรี พิริยะกุล. (2548). *ตัวแบบการทดแทนข้อมูลในการวิจัยทางสังคมศาสตร์ : การจำลองแบบกรณีตัวแบบอย่างง่าย*. กรุงเทพฯ: มหาวิทยาลัยรามคำแหง.

Acock, A.C. (2005). “Working With Missing Values,” *Journal of Marriage and Family*, 67(4), 1012 – 1028.

Adam, P. and M. Tshilidzi. (2009). “Evaluating the Impact of Missing Data Imputation,” *Advanced Data Mining and Applications*, 5678, 577 – 586. (Online). http://link.springer.com/chapter/10.1007%2F978-3-642-03348-3_59#page-1 [13 October 2011]

Andridge, R. R. and R. J. A. Little. (2010) “A Review of Hot Deck Imputation for Survey Non-Response,” *International Statistical Review*, 78(1), 40-64.

Chaimongkol, W. (2004). *Three Composite Imputation Methods for Item Nonresponse Estimation in Sample Surveys*. Thesis in Statistics Ph.D. Thailand: National Institute of Development Administration.

Chaimongkol, W. and P. Suwattee. (2004). *Weighted Nearest Neighbor and Regression Imputation*. 9th Asia-Pacific Decision Sciences Institute Conference. APDSI-KOPOMS Seoul. Korea. July 1-4, 2004.

Crawford, K.C. and others. (1995). *A Comparison of Differences in Meteorological Measurements Made by the Oklahoma Meso Network at Two Co-Located ASOS Sites*. Paper Presented at the Meeting of 11th Intermational Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology. Dallsa, TX.

Enders, C.K. (2001). "The Performance of the Full Information Maximum Likelihoods Estimator in Multiple Regression Models with Missing Data," *Educational and Psychological Measurement*, 61(5), 713 – 740.

Gabriel, L.S. and others. (2010). "Best Practices for Missing Data Management in Counseling Psychology," *Journal of Counseling Psychology*, 54(1), 1 – 10.

Granberg-Rademacker, J.S. (2007). "A Comparison of Three Approaches to Handling Incomplete State Level Data," *State Politics and Policy Quarterly,* 7(3), 325 – 338.

Guertin, W.H. (1968). "Comparison of Three Methods of Handling Missing Observations," *Psychological Reports*, 22(3), 896.

Hassani, B.T. and others. (2004). "Regeneration Imputation Models for Complex Stands of Southeastern British Columbia," *The Forestry Chronicle*, 80(2), 271 – 278.

Hedden, L.S. and others. (2008). "A Comparison of Missing Data methods for Hypothesis Tests of the Treatment Effect in Substance Abuse Clinical Trials : A Monte Carlo Simulation Study," *Substance Abuse Treatment, Prevention,* 3,13. (Online). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2441613/ [13 October 2011]

Huang, R. and K.C. Carriere. (2006). "Comparison of Methods for Incomplete Repeated Measures Data Analysis in Small Samples," *Statistical Planning and Inference*, 136(1), 235 – 247.

Jackson, E.C. (1968). "Missing Values in Linear Multiple Discriminate Analysis," *Biometric*, 24(4), 835 – 844.

James, W. (2010). "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data," *Public Opinion Quarterly*, 74(2), 223 – 243.

Kevin Strike et al. (2001). “Software Cost Estimation with Incomplete Data,” *IEEE Transactions on Software Engineering*, 27(10), 890-908.

Mark Huisman. (2000). “Imputation of Missing Item Responses: Some Simple Techniques,” *Quality and Quantity*, 34(4), 331–351.

Nikos Tsikriktsis. (2005). **“**A Review of Techniques for Treating Missing Data in OM Survey Research,” *Journal of Operations Management*, 24(1), 53-62.

O'Rourke, T. (2000). “DATA ANALYSIS: THE ART AND SCIENCE OF CODING AND ENTERING DATA,” *American Journal of Health Studies*, 16(3), 164-166**.**

Puma, J.M. (2009). *What to Do When Data are Missing in Group Randomized Controlled Trials*. Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.s. Department of Education.

Qinbao, S. and S. Martin. (2007). “A New Imputation Method for Small Software Project Data Sets,” *The Journal of Systems and Software*, 80(1), 51 – 62; January.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3), 537–560.

Roth, P.L. and others, (1996). “The Impact of Four Missing Data Techniques on Validity Estimates in Human Resource Management,” *Journal of Business and Psychology*, 11(1), 101 – 112.

Siripitayananon, P. (2002). *Data Mining Techniques for Handling a Missing Data Problem*. Ph.D. Thesis in Computer Science, Alabama : The University of Alabama.

Sudman, S. (1982). *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossy-Bass.

Sybil L. Crawford et al. (1995). “A Comparison of Analytic Methods for Non-random Missingness of Outcome Data**,”** *Journal of Clinical Epidemiology*, 48(2), 209-219.

Ting, H.L. (2010). “A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data,” *Quality and Quantiry*, 44(2), 277 – 287.

Velicer, W.F. and S.M. Colby. (2005).“A Comparison of Missing Data Procedures for Arima Time Series Analysis,” *Educational and Psychological Measurement,* 65(4), 596 – 615.

Viragoontavan, S. (2000). *Comparing Six Missing Data Methods within the Discriminant Analysis Context : A Monte Carlo Study*. Ph.D. Thesis in Education, U.S.A.: The Ohio State Universty.

Viviano, C.M.L. and G. Lovison, (2004). “A Comparison Between Simple and Multiple Imputation in Applying Logistic Regression Models,” *Roma : Italian Statistical Society*. (Online). http://www.sis- statistica.it/ files/pdf/atti/RSBa2004p339-342.pdf [13 October 2009]

Wood, A.M. and others. (2004). “Are Missing Outcome Data Adequately Handled?. A Review of Publish Randomized Controlled Trials in Major Medical Journals,” *Clinical Trial,* 1(4), 368-376,

Zhang, B. and C.M. Walker. (2008). “Impact of Missing Data on Person Model Fit and Person Trait Estimation,” *Applied Psychological Measurement*, 32(8), 466 – 479.

ภาคผนวก

ภาคผนวก ก
ชุดคำสั่งโดยใช้โปรแกรมภาษา อาร์ (R)

1. ชุดคำสั่งในการสุ่มตัวอย่าง

```
library(foreign)
x=read.spss("D:/data.sav")
population <- data.frame(x)
N <- nrow(population)
sample.rows.rep <- sample( 1:N, 500, replace = T)
table( sample.rows.rep)
sample.data.rep <- population[ sample.rows.rep , ]
sample.data.rep
write.csv( sample.data.rep, file = "d:/sample500.csv")
```

2. ชุดคำสั่งในการสร้างข้อมูลสูญหายแบบสุ่มสมบูรณ์

```
library(foreign)
for(i in 1:100){
d = read.csv("d:/sample500.csv")
X <- data.matrix(d)
N <-nrow(X)
pMiss <- 0.05
idMiss <- sample(1:N, N * pMiss)
nMiss <- length(idMiss)
m <- 5
howmanyMiss <- sapply(idMiss, function(x) sample(1:m, 1))
howmanyMiss
misscols<-lapply(howmanyMiss, function(x) sample(2:8, x))
for (ii in 1:nMiss){
        for (j in misscols[[ii]]){
                X[idMiss[ii],j]<-NA
        }
}
mydir <- format("d:/missing/n500m5/n500m5_")
myfile <- file.path(paste(mydir,i, ".csv"))
write.csv(X, file = myfile)
}
```

ภาคผนวก ข
การทดสอบความเป็นปกติของข้อมูล ความเป็นอิสระของข้อมูล

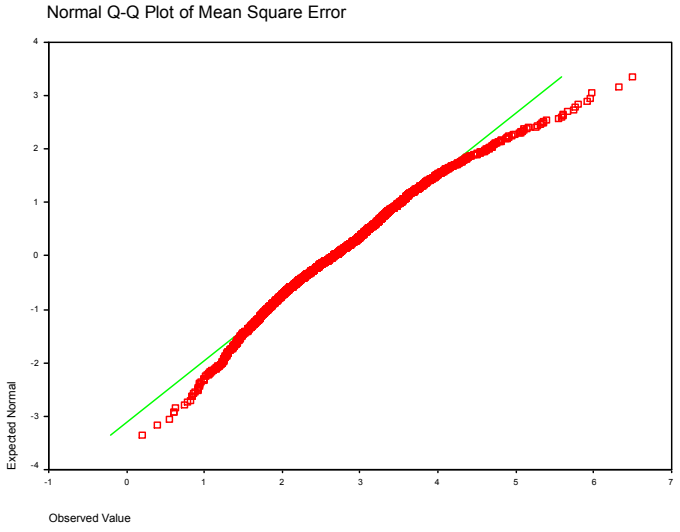
การทดสอบความเป็นปกติของข้อมูล
ความคลาดเคลื่อนกำลังสอง

Mean Square Error Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     1.00 Extremes    (=<.2)
      .00         0 .
     2.00         0 .  &
     5.00         0 .  6&
    17.00         0 .  899
    29.00         1 .  000111
    76.00         1 .  222222233333333
   110.00         1 .  444444444445555555555
   164.00         1 .  6666666666666777777777777777777
   194.00         1 .  888888888888888888899999999999999999999
   218.00         2 .  00000000000000000000000011111111111111111111
   212.00         2 .  22222222222222222222233333333333333333333
   199.00         2 .  44444444444444444444445555555555555555555
   222.00         2 .  66666666666666666666667777777777777777777777777
   212.00         2 .  88888888888888888888999999999999999999999999
   232.00         3 .  000000000000000000000000011111111111111111111111
   208.00         3 .  22222222222222222222223333333333333333333333333
   154.00         3 .  44444444444444455555555555555555555
   101.00         3 .  666666666667777777777
    75.00         3 .  888888889999999
    48.00         4 .  000001111
    36.00         4 .  2223333
    17.00         4 .  455
    22.00         4 .  66677
    13.00         4 .  889
     3.00         5 .  0
    25.00 Extremes    (>=5.1)

 Stem width:  1.000000
 Each leaf:       5 case(s)
 & denotes fractional leaves.
```



ภาพประกอบภาคผนวก **ข-1** กราฟ Normal q-q plot ของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

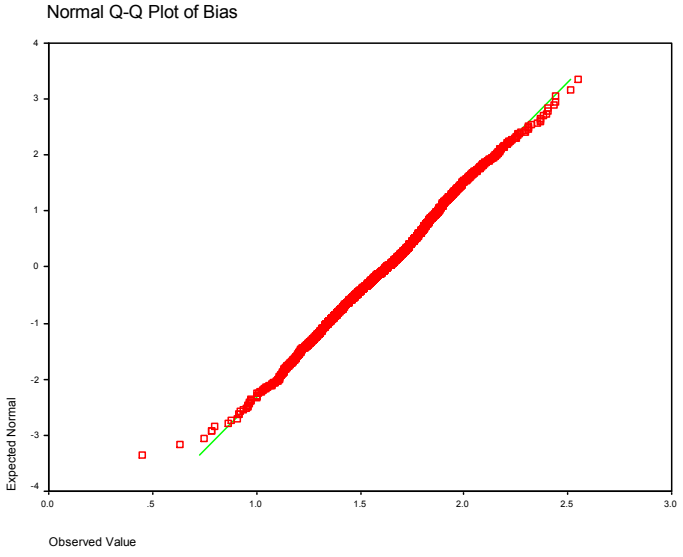
ความเอนเอียง

Bias Stem-and-Leaf Plot

Frequency Stem & Leaf

6.00 Extremes (=<.80)
2.00 8 . &
17.00 9 . &&
30.00 10 . 0&&
98.00 11 . 0123467899&
149.00 12 . 0011234456678899
264.00 13 . 000112223344455666778889999
304.00 14 . 0001112223334445566677778888899
326.00 15 . 000111222333344455566677778899
354.00 16 . 0001112222233344455566667778889999
385.00 17 . 00001112222233333444555566667778889999
330.00 18 . 0000011112222333444556667778889999
166.00 19 . 00122334455667899
85.00 20 . 0123456789
42.00 21 . 6&&&
19.00 22 . &&
7.00 23 . &
11.00 Extremes (>=2.37)

Stem width: .100000
Each leaf: 10 case(s)
& denotes fractional leaves.



ภาพประกอบภาคผนวก ข-2 กราฟ Normal q-q plot ของค่าความเอนเอียง

ความแปรปรวน

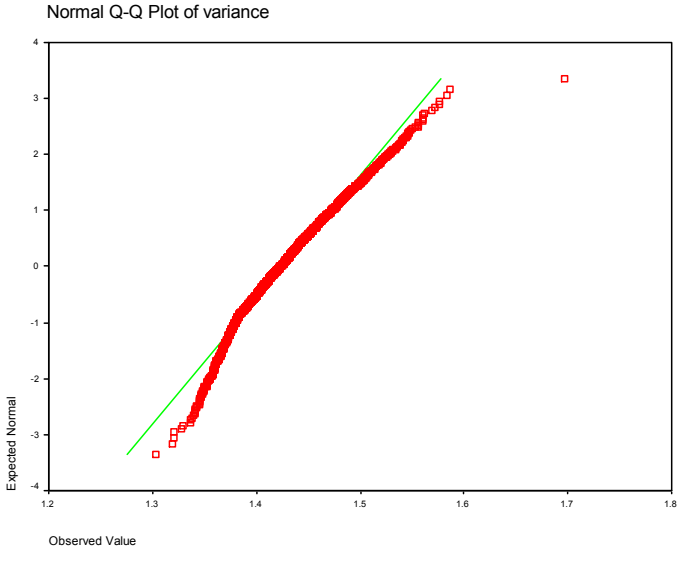
variance Stem-and-Leaf Plot

Frequency Stem & Leaf

1.00 130 . &
1.00 131 . &
4.00 132 . &
4.00 133 . &
29.00 134 . 5789&&
56.00 135 . 23458899&
126.00 136 . 00113344556667778899&
193.00 137 . 00111223333344455566667777888999
189.00 138 . 00001111122233344556667778899999
185.00 139 . 0001112223334444556667778889999
230.00 140 . 000011112222333344445555666677778889999
211.00 141 . 00011112222333334445566677778889999
222.00 142 . 000111222223334445556666777788889999
221.00 143 . 0001111222223333444555556667777888999

```
 181.00      144 .  000111112222333445556667788899
 172.00      145 .  000111223333444455677778888999
 126.00      146 .  00112222334455667889
 117.00      147 .  001112234566778889
  93.00      148 .  00122334567889
  58.00      149 .  012345789&
  58.00      150 .  11344567&&
  40.00      151 .  023469&
  28.00      152 .  29&&&
  18.00      153 .  &&&
  14.00      154 .  &&
  18.00 Extremes    (>=1.550)

Stem width:   .010000
Each leaf:       6 case(s)
& denotes fractional leaves.
```



ภาพประกอบภาคผนวก ข-3 กราฟ Normal q-q plot ของค่าความแปรปรวน

ความเป็นอิสระของข้อมูล

ตารางภาคผนวก ข-1 สถิติทดสอบความเป็นอิสระของข้อมูลระหว่างขนาดตัวอย่างกับวิธีการจัดการข้อมูลสูญหาย **Chi-Square Tests**

Statistics	Values					
	Value	df	Asymp. (2-sided)	Monte Carlo (2-sided)		
				p-value	99% Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	1.460	4	.834	.827	.818	.837
Likelihood Ratio	1.464	4	.833	.829	.819	.838
Fisher's Exact Test	1.457			.828	.819	.838
Linear-by-Linear Association	.001	1	.980	.983	.980	.987
N of Valid Cases	6971					

ตารางภาคผนวก ข-2 สถิติทดสอบความเป็นอิสระของข้อมูลระหว่างร้อยละข้อมูลสูญหายกับวิธีการจัดการข้อมูลสูญหาย **Chi-Square Tests**

Statistics	Value	df	Asymp. (2-sided)	Monte Carlo Sig. (2-sided)		
				p-value	99% Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	3.178(a)	8	.923	.928(b)	.921	.935
Likelihood Ratio	3.175	8	.923	.929(b)	.922	.935
Fisher's Exact Test	3.178			.928(b)	.921	.935
Linear-by-Linear Association	.298(c)	1	.585	.587(b)	.574	.600
N of Valid Cases	6971					

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.นิภาพร ชุติมันต์ ประธานกรรมการควบคุมวิทยานิพนธ์ อาจารย์ ดร.ประภาส ผิวอ่อน กรรมการควบคุมวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ทรงศักดิ์ ภูสีอ่อน ประธานกรรมการสอบ และผู้ช่วยศาสตราจารย์ ดร.สุนทรพจน์ ดำรงค์พานิช กรรมการสอบ

ขอขอบพระคุณคุณพ่อ คุณแม่ น้องชายที่คอยให้กำลังใจและอยู่เคียงข้างเสมอ คณาจารย์และเจ้าหน้าที่ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคามที่ให้คำแนะนำในการแก้ไขปัญหาต่าง ๆ ที่เกิดขึ้น

ไพฑูรย์ มุลิวัลย์

ชื่อเรื่อง การพัฒนาวิธีการประมาณค่าข้อมูลสูญหาย
ผู้วิจัย ไพฑูรย์ มุลิวัลย์
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** วิทยาการจัดการสถิติ
กรรมการควบคุม ผู้ช่วยศาสตราจารย์ ดร.นิภาพร ชุติมันต์
อาจารย์ ดร.ประภาส ผิวอ่อน
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2556

บทคัดย่อ

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี Hot-Deck Corrected Item Mean (HDD-CIM) และทำการเปรียบเทียบผลของการประมาณค่าข้อมูลสูญหายภายใต้กลไกการสูญหายของข้อมูลแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random-MCAR) และการสุ่มอย่างง่าย กับวิธีการประมาณค่าข้อมูลสูญหายแบบ Corrected Item Mean (CIM) และ Hot-Deck (HDD) โดยใช้ข้อมูลทุติยภูมิจากแฟ้มข้อมูลสำรวจความคิดเห็นของประชาชนเกี่ยวกับสถานการณ์การแพร่ระบาดยาเสพติด พ.ศ. 2554 ที่สำรวจโดยสำนักงานสถิติจังหวัดมหาสารคาม ที่มีขนาดตัวอย่าง 100 200 และ 500 ข้อมูลสูญหายร้อยละ 5 10 15 20 และ 30 ตามลำดับ

ผลการวิจัยปรากฏดังนี้

1. ผลการพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี HDD-CIM พบว่า วิธีการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ค่าความเอนเอียงค่าความแปรปรวนและค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ต่ำที่สุด นั่นแสดงว่า วิธีการประมาณค่าข้อมูลสูญหายที่พัฒนาขึ้นมีประสิทธิภาพในการประมาณค่าข้อมูลสูญหายดีกว่าวิธีที่นำมาเปรียบเทียบ

2. ผลการเปรียบเทียบประสิทธิภาพการประมาณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ความเอนเอียงและความแปรปรวน จำแนกตามขนาดตัวอย่าง ร้อยละของข้อมูลสูญหายและวิธีการการจัดการข้อมูลสูญหาย ที่ระดับนัยสำคัญ .05 สรุปผลได้ว่า ในทุก ๆ ขนาดตัวอย่าง การประมาณข้อมูลสูญหายแบบ HDD-CIM มีประสิทธิภาพสูงที่สุด

3. พบปฏิสัมพันธ์สองปัจจัยในการประมาณค่าความแปรปรวน 1) ปฏิสัมพันธ์ระหว่างขนาดตัวอย่าง วิธีการจัดการข้อมูลสูญหาย และ 2) ปฏิสัมพันธ์ระหว่างร้อยละข้อมูลสูญหาย วิธีการจัดการข้อมูลสูญหาย

โดยสรุป HDD-CIM มีประสิทธิภาพสูงสุดในทุก ๆ ขนาดตัวอย่างและทุกระดับร้อยละข้อมูลสูญหาย

คำสำคัญ : ข้อมูลสูญหาย, ข้อมูลสูญหายแบบสุ่มอย่างสมบูรณ์, Corrected Item Mean (CIM), วิธีการฮอทเดค (hot-deck)

TITLE The estimation development for missing data
AUTHOR Paitoon Muliwan
DEGREE Master Degree of Science **MAJOR** Statistical Management Science
ADVISORS Assist. Prof. Dr. Nipaporn Chutiman
Dr. Prapart Pue-on
UNIVERSITY Mahasarakham University **DATE** 2013

ABSTRACT

The purpose of this study was to develop the hot-deck corrected Item mean (HDD-CIM) for missing data estimation and compare its under missing complete at random (MCAR) and simple random sampling with two methods, namely; Corrected item mean (CIM) and hot-deck (HDD). The secondary data from a survey of public information about the prevalence of drugs, 2011, a survey by the Bureau of Statistics Mahasarakham Province were used and the comparisons were made with three sample sizes (100, 200 and 500) and four levels of percentage of missing data (5%, 10%, 15%, 20% and 30%). The results were as follows.

1. The development of a method for estimating the missing data with HDD-CIM found that mean square error, bias, variance and mean absolute percentage error had minimum values, its show that HDD-CIM was better way.

2. The performance comparison of estimation mean square error, bias and variance. By sample size, percentage of missing data and techniques for handing missing data. The .05 level of significance concluded that all sample sizes, the HDD-CIM has most powerful.

3. Two-way interactions were found : under the variance estimation i) sample size, techniques for handing missing data and ii) percentage of missing data, techniques for handing missing data.

In conclusion, hot deck corrected item mean method (HDD-CIM) performed most effectively with all sample size and all percentage of missing data.

Keywords : Missing data, MCAR, Simple random sampling, Corrected item mean (CIM), hot-deck (HDD)