บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัญหาหลักปัญหาหนึ่งจากการเก็บรวบรวมข้อมูลจากแบบสอบถามในการสำรวจตัวอย่างคือ การไม่ตอบ (nonresponse) หรือเรียกว่า ข้อมูลสูญหาย (missing data) ข้อมูลสูญหายจากการสำรวจตัวอย่างโดยทั่วไปมี 2 ประเภท คือ การไม่ตอบของหน่วยตัวอย่างบางหน่วย (unit nonresponse) และการไม่ตอบเฉพาะบางคำถามหรือบางตัวแปร (item nonresponse) โดย Kalton and Kasprzyk (1982) ได้ให้นิยามของการไม่ตอบของหน่วยตัวอย่างบางหน่วยและการไม่ตอบเฉพาะบางคำถามหรือบางตัวแปรไว้ว่า การไม่ตอบของหน่วยตัวอย่างบางหน่วย คือ การไม่ตอบสำหรับหน่วยตัวอย่างบางหน่วยซึ่งอาจเป็นผลสืบเนื่องมาจากการไม่เข้าใจความหมายของคำต่าง ๆ ที่ใช้ในแบบสอบถาม และการไม่ตอบเฉพาะบางคำถามหรือบางตัวแปร คือ การสูญหายของข้อมูลที่เกิดจากการไม่ตอบเฉพาะบางคำถามหรือบางตัวแปร ซึ่งอาจเกิดจากการออกแบบคำถามในแบบสอบถามไม่ครอบคลุมทำให้ผู้ตอบแบบสอบถามไม่สามารถตอบคำถามบางคำถามได้หรืออาจเกิดความผิดพลาดในการบันทึกข้อมูลทำให้ต้องตัดข้อมูลออกไป ส่วนใหญ่การแก้ไขปัญหาข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม วิธีการเชิงสถิติมักใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (imputation) แล้วนำค่าประมาณที่ได้ไปแทนค่าข้อมูลสูญหายเพื่อให้ได้ชุดข้อมูลสมบูรณ์สำหรับการวิเคราะห์เชิงสถิติต่อไป (Montaquila, 1998) ซึ่งในงานวิจัยนี้สนใจศึกษาข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามหรือบางตัวแปร

ข้อมูลสูญหายเป็นสถานการณ์ที่เกิดขึ้นจริงและไม่สามารถหลีกเลี่ยงได้ในงานวิจัยการตัดหน่วยตัวอย่างที่ไม่ตอบออกจากการวิเคราะห์ข้อมูลเป็นสิ่งที่อาจกระทำได้ถ้าข้อมูลนั้นหามาได้โดยง่ายมีค่าใช้จ่ายน้อยและมีการสำรองข้อมูลจำนวนมาก ๆ แต่ในความเป็นจริงแล้วอาจเสียเวลาค่าใช้จ่าย กำลังแรงงานและทรัพยากรอื่น ๆ เป็นจำนวนมาก ซึ่งวิธีการเชิงสถิติที่ใช้ในการประมาณค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณมีผู้ที่เสนอไว้หลายวิธี ได้แก่

1. วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean imputation หรือ *MI*)
2. วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย (Regression imputation หรือ *RI*)
3. วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยโดยสุ่ม (Random regression imputation หรือ *RRI*)
4. วิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด (Nearest neighbor imputation หรือ *NNI*)

5. วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก (Weighted nearest neighbor and regression imputation หรือ *WNR*)

6. วิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน (Ratio imputation หรือ *RAT*)

7. วิธีการประมาณข้อมูลสูญหายด้วยค่าฮอทเดค (Hot deck imputation หรือ *HDI*)

8. วิธีการประมาณข้อมูลสูญหายด้วยค่าโคลเดค (Cold deck imputation หรือ *CDI*)

9. วิธีการประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซ (Compromised imputationหรือ *COMP*)

10. วิธีการประมาณข้อมูลสูญหายของ Singh et al. (2010)

เป็นต้น

Chaimongkol and Suwattee (2005 c) ศึกษาวิจัยเรื่อง Weighted Nearest Neighbor and Regression Imputation ซึ่งเสนอวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก ซึ่งเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ได้นำวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุดและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยมารวมกันแล้วถ่วงน้ำหนัก ทำการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหาย 3 วิธี คือวิธีการประมาณข้อมูลสูญหายด้วยค่าใกล้สุด วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก พบว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนักเป็นวิธีที่เหมาะสมเมื่อพิจารณาค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean square error หรือ *RMSE*) ภายใต้วิธีการศึกษาจำลองด้วยวิธีมอนติคาร์โล (Monte Carlo method) และในปีเดียวกัน Chaimongkol and Suwattee ศึกษาวิจัยเรื่อง The Effect of Weighted Nearest Neighbor–Regression Imputation on Simple Linear Regression Analysis ซึ่งได้เปรียบเทียบประสิทธิภาพของตัวประมาณค่าเฉลี่ยประชากรในการวิเคราะห์การถดถอยเชิงเส้นเชิงเดียวเมื่อใช้วิธีการประมาณค่าข้อมูลสูญหาย 4 วิธี ได้แก่ วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก วิธีการประมาณข้อมูลสูญหายด้วยค่าฮอทเดค วิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยโดยสุ่ม พบว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนักเป็นวิธีที่เหมาะสมเมื่อพิจารณาค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean absolute deviation) และมีช่วงความเชื่อมั่นของค่าเฉลี่ยแคบกว่าวิธีอื่นภายใต้วิธีการศึกษาจำลองด้วยวิธีมอนติคาร์โล

Singh et al. (2010) แนะนำวิธีประมาณค่าข้อมูลสูญหายวิธีใหม่ โดยใช้เทคนิคการประมาณค่าในการสำรวจตัวอย่าง ซึ่งศึกษาประชากรจำกัดขนาด N เมื่อ X เป็นตัวแปรช่วยที่มีความสัมพันธ์กับตัวแปรที่สนใจ Y โดยการชักตัวอย่างแบบสุ่มเชิงเดียวชนิดไม่คืนที่ (Simple random sampling without replacement หรือ *SRSWOR*) ขนาด n จากประชากร N แล้วตัดข้อมูล

โดยสุ่ม $n-r$ หน่วย ใน n หน่วย กำหนดให้ R คือเซตของข้อมูลที่สมบูรณ์และ R^C คือเซตของข้อมูลสูญหาย ค่า y_i คือ ค่าข้อมูลสูญหายของหน่วยตัวอย่างที่ i และถูกประมาณค่าด้วยวิธีการประมาณค่าที่ใช้ค่าของข้อมูลของตัวแปรช่วย x ในเซตของข้อมูลที่สมบูรณ์และได้เสนอตัวประมาณค่าเฉลี่ยประชากร $\bar{Y}$ ภายใต้วิธีการประมาณค่าข้อมูลสูญหายดังกล่าว 2 ตัว คือ ตัวประมาณค่า T_{d1} และ T_{d2} และเปรียบเทียบประสิทธิภาพของตัวประมาณค่า T_{d1} และ T_{d2} กับตัวประมาณค่าเฉลี่ยประชากร เมื่อใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน วิธีการประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซ พบว่าตัวประมาณค่า T_{d1} มีประสิทธิภาพมากกว่าตัวประมาณค่าอื่น ๆ โดยศึกษาด้วยตัวแบบจำลอง

ดังนั้นในงานวิจัยนี้ ผู้วิจัยจึงมีความสนใจที่จะพัฒนาตัวประมาณค่า T_{d1} ให้มีประสิทธิภาพมากยิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อเสนอตัวประมาณค่าเฉลี่ยประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหายโดยการปรับตัวประมาณค่า T_{d1}

2. เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าที่นำเสนอกับตัวประมาณค่า T_{d1} และ T_{d2} ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีการประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน วิธีการประมาณข้อมูลสูญหายด้วยค่าคอมโพรไมซและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ได้แนวทางในการศึกษาและเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในการวิเคราะห์ข้อมูลรูปแบบอื่น ๆ ต่อไป

2. นำตัวประมาณค่าที่นำเสนอนี้ไปประยุกต์กับข้อมูลจริง

1.4 ขอบเขตของการวิจัย

1. การศึกษาครั้งนี้ศึกษาภายใต้วิธีการชักตัวอย่างแบบสุ่มเชิงเดียวชนิดไม่คืนที่เท่านั้น

2. การศึกษาเปรียบเทียบประสิทธิภาพของตัวประมาณค่าโดยใช้รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณค่า