บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

นักวิจัยที่ทำงานการวิจัยเชิงสำรวจจะพบกับปัญหาการไม่ตอบข้อคำถามบางข้อของกลุ่มตัวอย่าง เช่น คนที่อ่านหนังสือช้าจะไม่ตอบคำถามในตอนท้าย ๆ การสำรวจข้อมูลหลาย ๆ ครั้ง ผู้ให้ข้อมูลจะไม่ตอบคำถามในครั้งแรกของการสำรวจหรือบางครั้งจะไม่ตอบคำถามในการสำรวจซ้ำข้อมูลที่กลุ่มตัวอย่างไม่ดอบหรือขาดหายไปจะเรียกว่า ข้อมูลสูญหาย (missing data) การวัดซ้ำคุณลักษณะบางอย่างต้องสูญเสียทั้งเงินและเวลาจึงทำให้นักวิจัยเก็บข้อมูลซ้ำกับกลุ่มตัวอย่างบางคนเท่านั้น

ปัญหาการวิเคราะห์ข้อมูลที่มีข้อมูลสูญหายได้เกิดขึ้นมานานแล้วการแก้ปัญหาแบบที่พบโดยทั่วไปคือการตัดข้อมูลออกซึ่งเป็นวิธีที่ไม่ดีนัก การทำแบบนี้ทำให้กลุ่มตัวอย่างมีขนาดน้อยลงมีผลต่ออำนาจการทดสอบและการประมาณค่าพารามิเตอร์มีอคติ (Roth, 1994. pp. 538-539) การอ้างอิงข้อค้นพบที่ได้จากกลุ่มตัวอย่างไปยังประชากรมีความคลาดเคลื่อนสูง สิ่งที่สำคัญคือทำให้สูญเสียรายละเอียดบางอย่างไปซึ่งอาจจะมีผลกระทบต่อผลสรุปของการวิเคราะห์นั้น ๆ (วารุณี ตรีบำรุงศักดิ์, 2538. หน้า 2) และในการวิเคราะห์ข้อมูลโดยใช้โปรแกรมลิสเรลจะเกิดปัญหาสภาวะไม่เป็นบวกแน่นอน (Non - positive definite) ของเมตริกซ์ความแปรปรวนความแปรปรวนร่วมจากกลุ่มตัวอย่าง โปรแกรมคอมพิวเตอร์จะไม่แสดงผลอะไรเลย ผู้วิจัยจะต้องจัดการข้อมูลสูญหายด้วยวิธีใดวิธีหนึ่ง นอกจากการตัดข้อมูลสูญหายแบบแพร์ไวส์ (pairwise deletion) ก็จะสามารถวิเคราะห์ข้อมูลได้ (นงลักษณ์ วิรัชชัย, 2542. หน้า 307-308) ซึ่งในปัจจุบันนี้มีการแก้ปัญหาข้อมูลสูญหายโดยใช้วิธีการทางสถิติซึ่งสามารถแก้ปัญหาข้อมูลสูญหายได้เป็นอย่างดีและสามารถนำไปประยุกต์ใช้ในการทำวิจัยเพื่อปกป้องงานวิจัยไม่ให้เกิดการสรุปผลผิดพลาดจากการเกิดข้อมูลสูญหาย

ดังนั้นจึงได้มีการพัฒนาวิธีการจัดการข้อมูลสูญหายขึ้นมาเพื่อให้ข้อมูลที่สูญหายไปมีความสมบูรณ์ขึ้นก่อนที่จะนำไปวิเคราะห์ ลิทเทิลและรูบิน (Little & Rubin, 1987. pp. 6-7) ได้แบ่งวิธีการจัดการข้อมูลสูญหายไว้ 4 วิธี คือ

1. วิธีที่ใช้ข้อมูลสมบูรณ์ (Procedures based on completely recorded units) วิธีนี้เป็นวิธีที่ง่ายที่สุดโดยการตัดข้อมูลที่สูญหายออกไปแล้ววิเคราะห์ข้อมูลที่สมบูรณ์เท่านั้น เช่น การตัดข้อมูลสูญหายแบบลิสท์ไวส์ (listwise deletion) การตัดข้อมูลสูญหายแบบแพร์ไวส์ (pairwise deletion) เป็นวิธีที่มีอยู่ในโปรแกรมคอมพิวเตอร์ เช่น โปรแกรม SPSS

2. วิธีการแทนค่า (Imputation based procedures) วิธีนี้ใช้การแทนข้อมูลสูญหายด้วยค่าที่ได้จากวิธีการต่าง ๆ เช่น การแทนค่าแบบฮอทเดค (hot deck imputation) การแทนค่าโดยใช้ค่าเฉลี่ย (mean imputation) และการแทนค่าโดยวิธีการถดถอย (regression imputation)

3. วิธีการถ่วงน้ำหนัก (Weighting procedures) และ

4. วิธีการที่ได้จากการนิยามโมเดล (Model-based procedures) วิธีนี้ได้จากการนิยามโมเดลของข้อมูลสูญหายบางส่วนและใช้หลักการการอ้างอิงเกี่ยวกับไลค์ลิฮูด (likelihood) ในโมเดลด้วยวิธีการประมาณค่าพารามิเตอร์ที่เรียกว่าแม็กซิมัมไลค์ลิฮูด (maximum likelihood) โดยใช้วิธีการทำซ้ำ เช่น EM (expectation maximization) FIML (full information maximization likelihood estimation) MI (multiple imputation) เป็นต้น

วิธีการจัดการข้อมูลสูญหายที่นิยมใช้มีอยู่ 7 วิธี คือ

1. การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ (Listwise deletion) วิธีนี้จะตัดหน่วยวิเคราะห์ที่มีข้อมูลสูญหายออกไปนำข้อมูลที่สมบูรณ์เท่านั้นไปวิเคราะห์ ข้อดีของวิธีนี้คือเป็นวิธีที่ง่ายในการนำไปใช้ มีอยู่ในโปรแกรมสำเร็จรูปทั่วไป ปัญหาที่เกิดขึ้นจากการใช้วิธีนี้คือ ลดอำนาจของการทดสอบและความแปรปรวนของข้อมูลเพราะว่าขนาดของกลุ่มตัวอย่างน้อยลง ผลลัพท์ที่ได้จากการวิเคราะห์มีอคติ ค่าที่ได้ไม่ใช่ค่าที่แท้จริงของประชากรและไม่สามารถทราบสาเหตุของการสูญหายได้ ถึงแม้ว่าจะให้ค่าประมาณที่ไม่มีอคติเมื่อสาเหตุของการสูญหายเป็น missing completely at random แต่ก็ไม่แนะนำให้ใช้ด้วยข้อจำกัดดังกล่าว

2. การตัดข้อมูลสูญหายออกแบบแพร์ไวส์ (Pairwise deletion) วิธีนี้จะตัดหน่วยวิเคราะห์ที่มีข้อมูลสูญหายออกไปเมื่อนำข้อมูลของตัวแปรนั้นมาวิเคราะห์ ผลที่ได้ในการวิเคราะห์ความสัมพันธ์จะได้มาจากกลุ่มตัวอย่างที่แตกต่างกัน ถ้าข้อมูลมีการสูญหายแบบสุ่ม (random missing data) วิธีนี้ให้การประมาณค่าความสัมพันธ์ดีที่สุดเพราะใช้ข้อมูลทั้งหมดที่เก็บรวบรวมมา ปัญหาที่สำคัญ คือ เมตริกซ์สหสัมพันธ์จะไม่เป็นบวกแน่นอน (non-positive definite) ซึ่งจะพบได้ในการวิเคราะห์โดยใช้โปรแกรมลิสเรล (LISREL) ค่าความสัมพันธ์ที่ได้จากการตัดข้อมูลสูญหาย

แบบแพร์ไวส์จะมีค่ามากกว่าหรือน้อยกว่าค่าความสัมพันธ์ที่ได้จากกลุ่มตัวอย่างทั้งหมด วิธีนี้จะทำให้สูญเสียอำนาจการทดสอบน้อยกว่าวิธีลิสท์ไวส์ มีโปรแกรมสำเร็จรูปสำหรับวิเคราะห์ข้อมูลด้วยวิธีการตัดข้อมูลสูญหายแบบแพร์ไวส์

3. การแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean substitution) เป็นวิธีที่ใช้กันมากในการจัดการข้อมูลสูญหาย วิธีนี้จะแทนข้อมูลสูญหายด้วยค่าเฉลี่ยจากตัวแปรที่มีข้อมูลสมบูรณ์ กรณีที่ข้อมูลมีการแจกแจงเป็นโค้งปกติใช้ค่าเฉลี่ยจะเหมาะสม แต่ถ้าข้อมูลมีลักษณะการแจกแจงแบบเบ้ควรแทนค่าข้อมูลสูญหายด้วยมัธยฐานจะดีกว่า ถ้าข้อมูลมีการแจกแจงแบบปกติและเป็นข้อมูลสูญหายแบบสุ่มแล้วการแทนค่าด้วยวิธีนี้จะไม่มีอคติ แต่การแทนค่าแบบนี้ทุกคนที่ไม่ตอบจะถูกกำหนดให้มีคะแนนเท่ากันคือค่าเฉลี่ยซึ่งจริง ๆ แล้วน่าจะมีข้อมูลที่แตกต่างกัน ดังนั้นความแปรปรวนและความแปรปรวนร่วมระหว่างตัวแปรที่มีข้อมูลสูญหายจะมีค่าลดลง ซึ่งทำให้การประมาณค่า R^2 และ β น้อยลง

4. การแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของกลุ่มย่อย (Group mean substitution) วิธีนี้พัฒนามาจากวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย โดยการแทนค่าด้วยค่าเฉลี่ยจากกลุ่มตัวอย่างที่มีลักษณะคล้ายกับหน่วยข้อมูลที่มีข้อมูลสูญหาย ความแปรปรวนของตัวแปรที่มีข้อมูลสูญหายจะลดน้อยลงแต่ดีกว่าการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยเพราะข้อมูลสูญหายจะถูกแทนค่าด้วยค่าที่ไม่เหมือนกัน แม้กระนั้นก็ตามวิธีนี้ทำให้ความแปรปรวนและความแปรปรวนร่วมลดลงอย่างมีอคติแต่ก็ยังรักษาความแปรปรวนของตัวแปรที่มีข้อมูลสูญหายได้ดีกว่าวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ข้อดีก็คือแทนค่าข้อมูลสูญหายได้ถูกต้องมากกว่าแทนค่าด้วยค่าเฉลี่ยทั้งหมดเพราะว่าค่าที่ได้จะขึ้นอยู่กับลักษณะของกลุ่มย่อยที่มีตัวแปรอื่น ๆ เหมือนกันกับหน่วยข้อมูลที่มีข้อมูลสูญหาย แต่ถ้าลักษณะของตัวแปรแตกต่างจากหน่วยข้อมูลที่มีข้อมูลสูญหายก็ไม่สามารถแทนค่าข้อมูลสูญหายได้

5. การแทนค่าข้อมูลสูญหายโดยใช้วิธีการถดถอย (Regression imputation) วิธีนี้ใช้การวิเคราะห์การถดถอยเพื่อสร้างสมการทำนายข้อมูลสูญหายจากข้อมูลที่สมบูรณ์ทั้งหมด **สุนันทา วีรกุลเทวัญ (2544. หน้า 20) กล่าวว่า ถึงแม้ว่าวิธีการแทนค่าด้วยวิธีการถดถอยจะทำให้ความแปรปรวนของข้อมูลมีค่ามากกว่าความแปรปรวนที่ได้จากวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย แต่ก็มีข้อเสียคือจะทำให้ค่าที่เกี่ยวข้องกับความสัมพันธ์ (association) ระหว่างตัวแปรบิดเบือน (distort) ไปเพราะสมการประมาณค่าข้อมูลสูญหายสร้างมาจากความสัมพันธ์ระหว่างตัวแปร**

6. การแทนค่าข้อมูลสูญหายด้วยวิธีอีเอ็ม (EM algorithm) เป็นวิธีการแทนค่าข้อมูลสูญหายโดยใช้วิธีการถดถอยแล้วประมาณค่าเมตริกซ์ความแปรปรวนร่วม ดำเนินการทำซ้ำ (iterative method) จนค่าของเมตริกซ์ความแปรปรวนร่วมไม่แตกต่างกัน วิธีนี้มีความคลาดเคลื่อนแบบสุ่มรวมอยู่ในค่าที่ถูกทำนาย ถ้าสมการถดถอยให้ผลการทำนายที่ดีก็จะมีความคลาดเคลื่อนน้อย แต่ถ้าสมการถดถอยให้ผลการทำนายที่ไม่ดีจะมีความคลาดเคลื่อนมาก นักสถิติแนะนำให้ใช้ในการแก้ปัญหาข้อมูลสูญหายเพราะใช้วิธีการถดถอยซึ่งน่าจะแทนค่าข้อมูลสูญหายได้ถูกต้องมากยิ่งขึ้น

7. การแทนค่าข้อมูลสูญหายด้วยวิธีเอ็มไอ (Multiple imputation) วิธีนี้พัฒนามาจากวิธีอีเอ็ม (EM algorithm) แต่การทำซ้ำจะทำให้เกิดกลุ่มที่แตกต่างกันของค่าที่ถูกแทน ข้อมูลที่แตกต่างกันนำไปใช้ในการประมาณค่าความคลาดเคลื่อนในโมเดล วิธีนี้จะมีประโยชน์เพราะว่าการประมาณค่าความคลาดเคลื่อนมาตรฐานพิสูจน์ได้โดยวิธีการทางคณิตศาสตร์ แต่ขั้นตอนการทำงานต้องใช้เวลามากและต้องใช้โปรแกรมที่พัฒนาขึ้นมาสำหรับการวิเคราะห์เฉพาะ

จะเห็นว่าวิธีการจัดการข้อมูลสูญหายมีหลายวิธีทั้งวิธีธรรมดา เช่น การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ที่มีใช้อยู่ในโปรแกรมคอมพิวเตอร์ทั่วไป และวิธีที่ต้องใช้สถิติขั้นสูงประมาณค่าพารามิเตอร์ด้วยการทำซ้ำ เช่น วิธีการจัดการข้อมูลสูญหายแบบอีเอ็ม (EM algorithm) และวิธีการจัดการข้อมูลสูญหายแบบเอฟไอเอ็มแอล (FIML) เป็นวิธีการที่ดีและนำมาใช้ในงานวิจัยค่อนข้างมาก เช่น งานวิจัยของ ฟิงค์ไบเนอร์ (Finkbeiner, 1979. pp. 416-420) มาร์แคนโทนิโอ (Marcantonio, 1992. http://thailis.uni.net.th/dao/detail.nsp) และงานวิจัยที่ใช้การประมาณค่าด้วยการทำซ้ำลักษณะเช่นเดียวกับวิธีอีเอ็ม คือ งานวิจัยของ สุนันทา วีรกุลเทวัญ (Viragoontavan, 2000. http://thailis.uni.net.th/dao/detail.nsp) ใช้วิธีการแทนค่าข้อมูลสูญหายด้วยวิธีเอ็มไอ (Multiple imputation) โดยใช้โปรแกรม SOLAR และโปรแกรม NORM พบว่าวิธีการจัดการข้อมูลสูญหายแบบเอ็มไอมีประสิทธิภาพสูงสุด แต่วิธีการจัดการข้อมูลสูญหายแบบเอฟไอเอ็มแอล (FIML) และวิธีเอ็มไอ (Multiple imputation) ต้องใช้โปรแกรมที่เฉพาะเจาะจงและการคำนวณต้องใช้เวลานาน ดังนั้นวิธีการจัดการข้อมูลสูญหายแบบอีเอ็ม (EM algorithm) จึงได้ถูกนำมาใช้ในการจัดการข้อมูลสูญหายมากกว่าวิธีอื่น ๆ ถึงแม้ว่าวิธีการจัดการข้อมูลสูญหายแบบอีเอ็มจะเป็นวิธีการที่ดีและใช้กันอย่างแพร่หลายแต่ก็มีจุดบกพร่อง คือ การแทนค่าข้อมูลสูญหายครั้งแรกในสมการ $\hat{Y} = a + bX$ ใช้ค่าสถิติได้มาจากกลุ่มตัวอย่างที่มีข้อมูลสมบูรณ์ตัด

หน่วยตัวอย่างที่มีข้อมูลสูญหายออกไป การประมาณค่าพารามิเตอร์ด้วยค่าเหล่านี้จึงมีอคติ (bias) หมายความว่าค่าที่ได้อาจจะไม่ใช่ค่าที่แท้จริงของประชากร ดังนั้นผู้วิจัยจึงเสนอวิธีการประมาณค่าข้อมูลสูญหายโดยที่ขั้นตอนแรกประมาณค่าข้อมูลสูญหายด้วยวิธีการถดถอยอย่างง่าย ทำการประมาณค่าพารามิเตอร์ด้วยการทำซ้ำ นำค่าพารามิเตอร์ที่ได้ไปแทนในสมการ $\hat{Y} = a + bX$ แล้วจึงคัดเลือกสมการทำนายที่มีความคลาดเคลื่อนน้อยที่สุดจากการทำซ้ำ 1,000 ครั้ง เพื่อให้สมการทำนายข้อมูลสูญหายได้อย่างถูกต้องแม่นยำ เรียกวิธีนี้ว่า วิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี

การพิจารณาว่าวิธีการจัดการข้อมูลสูญหายวิธีใดดีกว่ากันหรือไม่ในการวิจัยนี้พิจารณาจากความแม่นยำ คือ ความใกล้เคียงกันระหว่างค่าสถิติกับค่าพารามิเตอร์ซึ่งก็คือความคลาดเคลื่อนถ้ามีความแม่นยำสูงค่าสถิติกับค่าพารามิเตอร์จะใกล้เคียงกันหรืออาจจะเป็นค่าเดียวกัน เป็นลักษณะของการศึกษาการประมาณค่าพารามิเตอร์แบบหนึ่งและนอกจากนั้นยังพิจารณาจากอำนาจการทดสอบว่าแตกต่างกันหรือไม่เพราะวิธีการจัดการข้อมูลสูญหายที่นิยมใช้คือการตัดข้อมูลสูญหายออกแบบลิสท์ไวส์นั้นทำให้กลุ่มตัวอย่างลดลงมีผลต่ออำนาจการทดสอบ สอดคล้องกับที่ ราจเมเคอร์ (Raaijmakers, 1999. p. 728) กล่าวว่า ประสิทธิภาพของวิธีการแทนค่าข้อมูลสูญหายจะต้องได้รับการประเมินจากการวิเคราะห์ทางสถิติที่หลากหลาย ดังได้กล่าวแล้วว่า ถ้ามีความแม่นยำสูงค่าสถิติกับค่าพารามิเตอร์จะใกล้เคียงกันหรือแตกต่างกันน้อยมาก เมื่อนำไปทดสอบสมมุติฐานทางสถิติโอกาสในการทำถูกก็น่าจะมีมาก นั่นก็คือ ถ้าสมมุติฐานสูญ (H_0) เป็นจริงก็น่าจะยอมรับสมมุติฐานสูญมาก แต่ถ้าสมมุติฐานสูญไม่จริงโอกาสในการปฏิเสธสมมุติฐานสูญก็น่าจะมีมาก การตัดสินใจปฏิเสธสมมุติฐานสูญที่ไม่จริงก็คือ อำนาจการทดสอบ ดังนั้นถ้าความแม่นยำสูง อำนาจการทดสอบก็น่าจะสูงด้วย แต่ถ้าความแม่นยำต่ำอำนาจการทดสอบก็น่าจะต่ำด้วย

ความแม่นยำและอำนาจการทดสอบจะขึ้นอยู่กับตัวแปรอื่น ๆ ด้วย ดังเช่นงานวิจัยของ เอ็นเดอร์ (Enders, 2001. pp. 713-740) ได้ศึกษาความสามารถของการประมาณค่าแบบเอฟไอเอ็มแอล (Full information maximum likelihood : FIML) ในการวิเคราะห์การถดถอยพหุคูณเมื่อมีข้อมูลสูญหาย ตัวแปรที่ศึกษามี 4 ตัวแปร คือ วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย ขนาดของกลุ่มตัวอย่าง และขนาดของความสัมพันธ์ระหว่างตัวแปร พบว่าวิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และขนาดความสัมพันธ์ระหว่างตัวแปร

มีปฏิสัมพันธ์กันและเมื่อผู้วิจัยทำการวิจัยเชิงสำรวจกับประชากรที่มีขนาดใหญ่ ผู้วิจัยจำเป็นต้องสุ่มตัวอย่างขึ้นมาศึกษา ดังนั้นความแม่นยำและอำนาจการทดสอบน่าจะขึ้นอยู่กับตัวแปรวิธีการสุ่มตัวอย่างด้วย ซึ่งสอดคล้องกับคำกล่าวของ คอมเรย์และฮินส์ (Kromrey & Hines, 1994. p. 575) ที่กล่าวว่า การศึกษาวิธีการจัดการข้อมูลสูญหายจะต้องศึกษาว่าเมื่อใช้การสุ่มที่แตกต่างกันประมาณค่าพารามิเตอร์ได้แตกต่างกันหรือไม่ และงานวิจัยของ ราจเมเคอร์ (Raaijmakers, 1999. p. 728) ที่พบว่า ความแตกต่างของวิธีการจัดการข้อมูลสูญหายจะลดลงด้วยปัจจัยต่อไปนี้ ขนาดของกลุ่มตัวอย่างมากขึ้น จำนวนเปอร์เซ็นต์ของการสูญหายน้อย ตัวแปรสูญหายน้อย และการลดลงในระดับความสัมพันธ์ระหว่างตัวแปร จากการศึกษางานวิจัยเกี่ยวกับการสุ่มตัวอย่าง พบว่า วิธีการสุ่มตัวอย่างต่างกันประมาณค่าพารามิเตอร์ได้ต่างกันโดยเฉพาะการสุ่มแบบแบ่งชั้นที่ใช้ตัวแปรจำแนกชั้นภูมิเป็นขนาดโรงเรียน กำหนดขนาดกลุ่มตัวอย่างย่อยแบบนีย์แมนใช้การสุ่มแบบมีระบบภายในชั้นภูมิทำให้การประมาณค่าพารามิเตอร์มีประสิทธิภาพสูงสุด แต่ถ้าใช้การสุ่มแบบกลุ่มจะต้องใช้การสุ่มตัวอย่างย่อยแบบง่ายจะทำให้การประมาณค่าพารามิเตอร์ดีกว่าวิธีการสุ่มตัวอย่างย่อยแบบมีระบบ (สมชัย วงษ์นายะ, 2533 ; ดวงใจ ปวีณอภิชาต, 2535 ; สุกัญญรัตน์ คงงาม, 2539)

ลักษณะความสัมพันธ์ระหว่างตัวแปรที่ผู้วิจัยสุ่มตัวอย่างขึ้นมาศึกษาจะแตกต่างกันไปตามตัวแปรที่สนใจศึกษา รูท (Roth, 1994. pp. 542-543) ได้ศึกษาพบว่า ในกรณีที่ความสัมพันธ์ระหว่างตัวแปรมีค่าสูงวิธีการจัดการข้อมูลสูญหายโดยการถดถอยจะดีกว่าความสัมพันธ์ระหว่างตัวแปรที่มีค่าต่ำ (r=.20 ถึง .30 หรือต่ำกว่านี้) แต่ทั้งนี้ก็ขึ้นอยู่กับตัวแปรอื่น ๆ ด้วย ซึ่งข้อค้นพบขัดแย้งกับงานวิจัยของ ราจเมเคอร์ (Raaijmakers, 1999. p. 728) ที่ได้กล่าวว่า วิธีการจัดการข้อมูลสูญหายมีหลายวิธีความแตกต่างของวิธีการจะน้อยลงจากปัจจัยต่าง ๆ ต่อไปนี้ ขนาดของกลุ่มตัวอย่างมากขึ้น จำนวนข้อมูลสูญหายน้อย ตัวแปรที่มีข้อมูลสูญหายน้อย และความสัมพันธ์ระหว่างตัวแปรน้อยลง

เมื่อผู้วิจัยดำเนินการเก็บรวบรวมข้อมูลมาแล้วจะมีจำนวนข้อมูลสูญหายแตกต่างกันขึ้นอยู่กับข้อคำถามและความตั้งใจของผู้ตอบ ถ้ามีข้อมูลสูญหายเป็นจำนวนน้อย เช่น ประมาณ 5% การตัดหน่วยตัวอย่างออกไปดูเหมือนว่าจะมีเหตุผลในการแก้ปัญหาข้อมูลสูญหาย แต่ถ้ามีข้อมูลสูญหายจำนวนมากการตัดข้อมูลออกไปจะไม่มีประสิทธิภาพข้อมูลที่เหลืออยู่จะไม่เป็นตัวแทนของประชากรซึ่งมีเป้าหมายในการอ้างอิง (Schafer, 1997. p. 1 ; Little & Rubin, 1987. p. 5) แต่ถ้าใช้วิธีการถดถอยจัดการข้อมูลสูญหายจะมีความเหมาะสมเมื่อมีข้อมูลสูญหาย

มากกว่า 20% (Roth, 1994. p. 542 citing Raymond & Roberts, 1987) การเลือกใช้วิธีการจัดการข้อมูลสูญหายควรพิจารณาองค์ประกอบอื่น ๆ ด้วย องค์ประกอบที่สำคัญ คือ จำนวนข้อมูลสูญหาย (Roth, 1994. p. 550)

จากจุดอ่อนของวิธีการจัดการข้อมูลสูญหายแบบอีเอ็ม (EM algorithm) ประกอบกับมีตัวแปรที่เกี่ยวข้องกับการศึกษาวิธีการจัดการข้อมูลสูญหายหลายตัวแปร ผู้วิจัยจึงสนใจที่จะพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี (EPSSE) และตรวจสอบความแม่นยำและอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายที่พัฒนาขึ้นกับวิธีอีเอ็ม (EM algorithm) ซึ่งเป็นวิธีการจัดการข้อมูลสูญหายที่นิยมใช้กันมากการประมาณค่าพารามิเตอร์ด้วยการทำซ้ำทำให้มีความถูกต้องและแม่นยำสูงและใช้วิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์มาเป็นพื้นฐานในการเปรียบเทียบ เพราะวิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์จะตัดหน่วยตัวอย่างที่มีข้อมูลสูญหายออกไปจากการวิเคราะห์เป็นวิธีที่ทำแล้วข้อมูลที่เหลืออยู่เป็นข้อมูลจริง การนำไปเปรียบเทียบกับวิธีการจัดการข้อมูลสูญหายแบบอื่น ๆ ทำให้เห็นความแตกต่างได้อย่างชัดเจนว่าการตัดออกไปหรือการแทนค่าวิธีใดดีกว่ากัน โดยใช้การศึกษาแบบการจำลองสถานการณ์ เทคนิคมอนติคาร์โล ซิมูเลชั่น (Monte Carlo Simulation) จะทำให้ได้ผลการศึกษาที่แน่นอน เพราะสามารถกำหนดสถานการณ์ต่าง ๆ ได้ครอบคลุมกับสถานการณ์จริงที่จะเกิดขึ้น

ปัญหาการวิจัย

1. วิธีการจัดการข้อมูลสูญหายที่พัฒนาขึ้นมีความแม่นยำและอำนาจการทดสอบเป็นอย่างไร

2. มีปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ระหว่างตัวแปร ที่ส่งผลต่อความแม่นยำหรือไม่

วัตถุประสงค์ของการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์ของการวิจัยดังต่อไปนี้

1. **เพื่อพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี**

2. เพื่อตรวจสอบความแม่นยำและอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน โดยพิจารณาเปรียบเทียบดังนี้

2.1 เพื่อเปรียบเทียบความแม่นยำของค่าเฉลี่ยเลขคณิตที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2.2 เพื่อเปรียบเทียบความแม่นยำของความแปรปรวนที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2.3 เพื่อเปรียบเทียบความแม่นยำของค่าสัมประสิทธิ์สหสัมพันธ์ที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2.4 เพื่อเปรียบเทียบอำนาจการทดสอบของข้อมูลที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

3. เพื่อศึกษาปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ระหว่างตัวแปร ที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์

สมมุติฐานการวิจัย

การวิจัยครั้งนี้ผู้วิจัยต้องการพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี และตรวจสอบความแม่นยำและอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี แบบอีเอ็ม และแบบลิสท์ไวส์ และจากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า การศึกษาวิธีการจัดการข้อมูลสูญหายจะมีตัวแปรที่เข้ามาเกี่ยวข้อง คือ วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหาย ดังนั้นผู้วิจัยจึงตั้งสมมุติฐานการวิจัยดังนี้

1. วิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี น่าจะมีความแม่นยำและอำนาจการทดสอบดีกว่าวิธีอื่นเมื่อใช้วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2. ความแม่นยำน่าจะขึ้นอยู่กับปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ระหว่างตัวแปรที่แตกต่างกัน

ขอบเขตของการวิจัย

ขอบเขตของการวิจัยในครั้งนี้มีดังต่อไปนี้

1. **ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลที่ได้จากการจำลองสถานการณ์โดยใช้เทคนิค** มอนติ คาร์โล ซิมูเลชั่น (Monte Carlo Method)

2. ความแม่นยำ พิจารณาจาก ความใกล้เคียงกันระหว่างค่าสถิติกับค่าพารามิเตอร์ พิจารณาเฉพาะค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ ทั้งนี้ที่พิจารณาดังกล่าวนี้เนื่องมาจาก ความแม่นยำที่ใช้ในการศึกษาก็คือ ความคลาดเคลื่อน เป็นลักษณะการศึกษาค่าประมาณพารามิเตอร์แบบหนึ่ง ถ้ามีความแม่นยำสูงค่าสถิติกับค่าพารามิเตอร์ใกล้เคียงกันมากการประมาณค่าพารามิเตอร์ย่อมมีความถูกต้อง แต่ถ้าความแม่นยำต่ำค่าสถิติกับค่าพารามิเตอร์จะแตกต่างกันมาก การประมาณค่าพารามิเตอร์จะมีความผิดพลาด

3. ตัวแปรที่ศึกษามีดังต่อไปนี้

3.1 ตัวแปรอิสระ

3.1.1 วิธีการสุ่ม จำแนกเป็น 3 วิธี คือ

3.1.1.1 การสุ่มแบบแบ่งชั้น

3.1.1.2 การสุ่มแบบกลุ่ม

3.1.1.3 การสุ่มแบบหลายขั้นตอน

3.1.2 วิธีการจัดการข้อมูลสูญหาย จำแนกเป็น 3 วิธี คือ

3.1.2.1 การตัดข้อมูลออกแบบลิสท์ไวส์ (Listwise deletion)

3.1.2.2 การแทนค่าข้อมูลด้วยวิธีอีเอ็ม (Em algorithm or Expectation maximization)

3.1.2.3 การแทนค่าข้อมูลด้วยวิธีอีพีเอสเอสอี (Estimated parameter and Smallest standard error)

3.1.3 ความสัมพันธ์ระหว่างตัวแปร 3 ลักษณะ คือ ความสัมพันธ์ระดับต่ำ (r=.30) ความสัมพันธ์ระดับปานกลาง (r=.50) และความสัมพันธ์ระดับสูง (r=.70)

3.1.4 จำนวนข้อมูลสูญหาย 5% 10% 20% 30%

3.2 ตัวแปรตาม คือ ความแม่นยำ และอำนาจการทดสอบ

4. อำนาจการทดสอบคำนวณจากการใช้สถิติทดสอบที (t-test) การวิเคราะห์ความแปรปรวนและการทดสอบสหสัมพันธ์

5. การวิจัยครั้งนี้ใช้ข้อมูลที่มีลักษณะการแจกแจงแบบปกติสองตัวแปร (Bivariate normal distribution) โดยกำหนดให้ตัวแปรเกณฑ์ (Y) คือ เกรดเฉลี่ยของนักเรียน และตัวแปรทำนาย (X) คือ คะแนนสอบของนักเรียน และกำหนดให้มีความสัมพันธ์ต่ำ (r=.30) ปานกลาง (r=.50) และสูง (r=.70)

6. ข้อมูลสูญหายที่จะศึกษาในครั้งนี้จะศึกษาเฉพาะตัวแปรเกณฑ์ (Y) คือ เกรดเฉลี่ยของนักเรียน

7. การประมาณขนาดตัวอย่างเป็นไปตามวิธีของนีย์แมน

ข้อตกลงเบื้องต้น

การดำเนินการสุ่มตัวอย่างข้อมูลที่ได้จากการจำลองสถานการณ์โดยใช้เทคนิคมอนติคาร์โล ซิมูเลชั่น (Monte Carlo Simulation Technique) กระทำการสุ่มซ้ำวิธีละ 1,000 ครั้ง ซึ่งมีความเพียงพอที่จะศึกษาค่าประมาณพารามิเตอร์

คำจำกัดความที่ใช้ในการวิจัย

ข้อมูลสูญหาย หมายถึง การที่พลวิจัยหน่วยนั้นไม่มีข้อมูลของการได้รับการสำรวจที่ครบถ้วนโดยที่ในการวิจัยนี้จำลองข้อมูลคะแนนสอบและเกรดเฉลี่ยขึ้นมาแล้วสุ่มข้อมูลจากตัวแปรตามซึ่งเป็นเกรดเฉลี่ยให้หายไป

ความแม่นยำ หมายถึง ความใกล้เคียงกันของค่าสถิติกับค่าพารามิเตอร์ โดยใช้เกณฑ์ตัดสินความแม่นยำ คือ ค่าเฉลี่ยกำลังสองของความแม่นยำมีค่าน้อยที่สุด

ความแม่นยำสูง หมายถึง ค่าเฉลี่ยกำลังสองของความแม่นยำมีค่าน้อย

ความแม่นยำต่ำ หมายถึง ค่าเฉลี่ยกำลังสองของความแม่นยำมีค่ามาก

อำนาจการทดสอบ หมายถึง ความน่าจะเป็นที่ปฏิเสธสมมุติฐานสูญ เมื่อสมมุติฐานนั้นไม่เป็นความจริง

วิธีการจัดการข้อมูลสูญหาย หมายถึง การกระทำเพื่อให้ข้อมูลที่สูญหายไปมีความสมบูรณ์ขึ้นก่อนที่จะนำไปวิเคราะห์ การวิจัยนี้กำหนดวิธีการจัดการข้อมูลสูญหาย 3 วิธี คือ การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ การแทนค่าข้อมูลสูญหายด้วยวิธีอีเอ็ม การแทนค่าข้อมูลสูญหายด้วยวิธีอีพีเอสเอสอี

การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ (Listwise deletion) หมายถึง การตัดพลวิจัยที่มีรายการคำตอบใดคำตอบหนึ่งเป็นข้อมูลสูญหายออกไปจากการวิเคราะห์

การแทนค่าข้อมูลสูญหายด้วยวิธีอีเอ็ม (EM algorithm) หมายถึง การแทนค่าข้อมูลสูญหายโดยใช้วิธีการถดถอยแล้วประมาณค่าเมตริกซ์ความแปรปรวนร่วม ดำเนินการทำซ้ำ (iterative method) จนค่าของเมตริกซ์ความแปรปรวนร่วมไม่แตกต่างกัน

การแทนค่าข้อมูลด้วยวิธีอีพีเอสเอสอี (Estimated parameter and the smallest standard error : EPSSE) หมายถึง การแทนค่าข้อมูลสูญหายโดยการประมาณค่าพารามิเตอร์ของข้อมูล สร้างสมการถดถอยทำนายค่าของข้อมูลสูญหาย แล้วคัดเลือกสมการถดถอยที่มีความคลาดเคลื่อนน้อยที่สุดในการทำนายข้อมูลสูญหายขั้นตอนสุดท้าย

วิธีการสุ่มตัวอย่าง หมายถึง การสุ่มแบบอิงความน่าจะเป็นซึ่งในการวิจัยครั้งนี้ใช้วิธีการสุ่ม 3 วิธี คือ การสุ่มแบบกลุ่ม การสุ่มแบบแบ่งชั้น และการสุ่มแบบหลายขั้นตอน

วิธีการสุ่มแบบแบ่งชั้น หมายถึง การสุ่มตัวอย่างโดยอาศัยหลักความน่าจะเป็น มีขั้นตอนการสุ่มดังนี้

1. แบ่งประชากรออกเป็นระดับชั้น 3 ระดับชั้น
2. กำหนดขนาดกลุ่มตัวอย่างในแต่ละระดับชั้นแบบนีย์แมน
3. สุ่มตัวอย่างจากประชากรที่ได้จำแนกไว้ในข้อที่ 1 ด้วยขนาดที่กำหนดไว้ในแต่ละระดับชั้นในข้อที่ 2 โดยใช้วิธีการสุ่มแบบเป็นระบบ

วิธีการสุ่มแบบกลุ่ม หมายถึง การสุ่มตัวอย่างโดยอาศัยหลักความน่าจะเป็น มีขั้นตอนการสุ่มดังนี้

1. แบ่งประชากรออกเป็นกลุ่มทั้งหมด 14 กลุ่ม
2. สุ่มกลุ่มออกมาครึ่งหนึ่งโดยการสุ่มอย่างง่าย
3. กำหนดขนาดกลุ่มตัวอย่างในแต่ละกลุ่มแบบนีย์แมน
4. สุ่มตัวอย่างตามกลุ่มที่สุ่มได้ในข้อที่ 2 ตามขนาดที่กำหนดไว้ในข้อที่ 3 โดยการสุ่มอย่างง่าย

วิธีการสุ่มแบบหลายขั้นตอน หมายถึง การสุ่มตัวอย่างโดยอาศัยหลักความน่าจะเป็น มีขั้นตอนการสุ่มดังนี้

1. แบ่งประชากรออกเป็นกลุ่มทั้งหมด 14 กลุ่ม
2. สุ่มกลุ่มออกมาครึ่งหนึ่งโดยการสุ่มอย่างง่าย

3. ในแต่ละกลุ่มแบ่งประชากรออกเป็นระดับชั้น 3 ระดับชั้น

4. กำหนดขนาดกลุ่มตัวอย่างในแต่ละระดับชั้นแบบนีย์แมน

5. สุ่มตัวอย่างจากประชากรที่จำแนกไว้ในข้อที่ 3 ด้วยขนาดที่กำหนดไว้ในข้อที่ 4 โดยวิธีการสุ่มแบบเป็นระบบ

ค่าพารามิเตอร์ หมายถึง ค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์
คำนวณจากข้อมูลทั้งหมดซึ่งคือประชากรที่ได้จากการจำลองสถานการณ์

ค่าสถิติ หมายถึง ค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ คำนวณจากกลุ่มตัวอย่างข้อมูลที่ได้จากการจำลองสถานการณ์ โดยใช้วิธีการสุ่ม วิธีการจัดการข้อมูลสูญหาย ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

ปฏิสัมพันธ์ หมายถึง ผลต่างที่ไม่เท่ากันของความแม่นยำภายใต้ระดับหนึ่ง ๆ ของตัวแปรหนึ่งเมื่อเปรียบเทียบไปยังอีกระดับหนึ่งของตัวแปรหนึ่ง

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ทำให้ได้วิธีการจัดการสูญหายแบบอีพีเอสเอสอี

2. ทำให้ทราบว่าค่าสถิติต่าง ๆ ที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายที่ต่างกันวิธีใดได้ค่าใกล้เคียงหรือเท่ากันกับค่าพารามิเตอร์

3. ได้องค์ความรู้เกี่ยวกับวิธีการจัดการข้อมูลสูญหาย

4. เป็นแนวทางสำหรับผู้วิจัยในการเลือกกลุ่มตัวอย่าง วิธีการจัดการข้อมูลที่สูญหาย ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายให้เหมาะสมกับสภาพของการวิจัยต่อไป