# การพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี และการตรวจสอบความแม่นยำและอำนาจการทดสอบเปรียบเทียบกับวิธีอีเอ็มและลิสท์ไวส์ : เทคนิคมอนติคาร์โล[1]

เชาว์ อินไชย[2] และ ดร.รัตนะ บัวสนธ์[3]

**บท**ความนี้นำเสนอวิธีการจัดการข้อมูลสูญหายแบบใหม่ คือ วิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีและตรวจสอบความแม่นยำและอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี แบบอีเอ็มและแบบลิสท์ไวส์ วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่มและแบบหลายขั้นตอน ความสัมพันธ์ของตัวแปรระดับต่ำ ปานกลางและสูง และจำนวนข้อมูลสูญหาย 5% 10% 20% และ 30% และศึกษาปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย ความสัมพันธ์ของตัวแปรและจำนวนข้อมูลสูญหาย ที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ข้อมูล ที่ใช้มีลักษณะการแจกแจงแบบปกติสองตัวแปร และใช้เทคนิคมอนติ คาร์โล ซิมูเลชั่น จำลองการทดลองด้วยเครื่องคอมพิวเตอร์โดยการทำซ้ำจำนวน 1,000 ครั้ง ผลการวิจัยพบว่าวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีได้ค่าความแม่นยำของค่าเฉลี่ยเลขคณิตไม่แตกต่างจากวิธีอีเอ็มและเป็นค่าที่สูงที่สุด วิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์ได้ค่าความแม่นยำของความแปรปรวนและค่าความแม่นยำของสัมประสิทธิ์สหสัมพันธ์สูงที่สุด วิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีได้ค่าอำนาจการทดสอบเมื่อใช้การทดสอบความสัมพันธ์สูงกว่าวิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์และแบบอีเอ็ม เมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น จำนวนข้อมูลสูญหายทุกระดับและความสัมพันธ์ของตัวแปรอยู่ในระดับต่ำและไม่มีปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย ความสัมพันธ์ของตัวแปรและจำนวนข้อมูลสูญหายที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ที่ระดับนัยสำคัญ .01

---

[1] วิทยานิพนธ์ ระดับดุษฎีบัณฑิต สาขาวิจัยและประเมินผลการศึกษา มหาวิทยาลัยนเรศวร

[2] ดุษฎีบัณฑิต สาขาวิจัยและประเมินผลการศึกษา มหาวิทยาลัยนเรศวร

[3] รองศาสตราจารย์ ประจำภาควิชาการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยนเรศวร

# บทนำ

นักวิจัยที่ทำงานการวิจัยเชิงสำรวจจะพบกับปัญหาการไม่ตอบข้อคำถามบางข้อของกลุ่มตัวอย่าง ข้อมูลที่กลุ่มตัวอย่างไม่ตอบหรือขาดหายไปจะเรียกว่า ข้อมูลสูญหาย (Missing data) การแก้ปัญหาที่พบโดยทั่วไปคือการตัดข้อมูลออกซึ่งเป็นวิธีที่ไม่ดีนัก การทำแบบนี้ทำให้กลุ่มตัวอย่างมีขนาดน้อยลงมีผลต่ออำนาจการทดสอบและการประมาณค่าพารามิเตอร์มีอคติ (Roth, 1994, pp. 538-539) ในปัจจุบันมีการแก้ปัญหาข้อมูลสูญหาย โดยใช้วิธีการทางสถิติซึ่งสามารถแก้ปัญหาข้อมูลสูญหายได้เป็นอย่างดีและสามารถนำไปประยุกต์ใช้ในการทำวิจัยเพื่อปกป้องงานวิจัยไม่ให้เกิดการสรุปผลผิดพลาดจากการเกิดข้อมูลสูญหาย

วิธีการจัดการข้อมูลสูญหายมีหลายวิธีทั้งวิธีธรรมดา เช่น การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ที่มีใช้อยู่ในโปรแกรมคอมพิวเตอร์ทั่วไป และวิธีที่ต้องใช้สถิติขั้นสูงประมาณค่าพารามิเตอร์ด้วยการทำซ้ำ เช่น วิธีการจัดการข้อมูลสูญหายแบบอีเอ็ม (EM algorithm) วิธีการจัดการข้อมูลสูญหายแบบเอฟไอเอ็มแอล (FIML) และวิธีการจัดการข้อมูลสูญหายแบบเอ็มไอ (Multiple imputation) แต่วิธีการจัดการข้อมูลสูญหายแบบเอฟไอเอ็มแอล (FIML) และแบบเอ็มไอ (Multiple imputation) ต้องใช้โปรแกรมที่เฉพาะเจาะจงและการคำนวณต้องใช้เวลานาน ดังนั้นวิธีการจัดการข้อมูลสูญหายแบบอีเอ็ม (EM algorithm) จึงได้ถูกนำมาใช้ในการจัดการข้อมูลสูญหายมากกว่าวิธีอื่น ๆ

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

จากจุดอ่อนของวิธีการจัดการข้อมูลสูญหายแบบอีเอ็ม (EM algorithm) ประกอบกับมีตัวแปรที่เกี่ยวข้องกับการศึกษาวิธีการจัดการข้อมูลสูญหายหลายตัวแปร ผู้วิจัยจึงสนใจที่จะพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี (EPSSE) และตรวจสอบความแม่นยำและอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายที่พัฒนาขึ้นกับวิธีอีเอ็ม (EM algorithm) ซึ่งเป็นวิธีการจัดการข้อมูลสูญหายที่นิยมใช้กันมากการประมาณค่าพารามิเตอร์ด้วยการทำซ้ำทำให้มีความถูกต้องและแม่นยำสูงและใช้วิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์มาเป็นพื้นฐานในการเปรียบเทียบ เพราะวิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์จะตัดหน่วยตัวอย่างที่มีข้อมูลสูญหายออกไปจากการวิเคราะห์เป็นวิธีที่ทำแล้วข้อมูลที่เหลืออยู่เป็นข้อมูลจริง การนำไปเปรียบเทียบกับวิธีการจัดการข้อมูลสูญหายแบบอื่น ๆ ทำให้เห็นความแตกต่างได้อย่าง

ชัดเจนว่าการตัดออกไปหรือการแทนค่าวิธีใดดีกว่ากัน โดยใช้การศึกษาแบบการจำลองสถานการณ์ เทคนิคมอนติ คาร์โล ชิมูเลชั่น (Monte Carlo Simulation) จะทำให้ได้ผลการศึกษาที่แน่นอน เพราะสามารถกำหนดสถานการณ์ต่าง ๆ ได้ครอบคลุมกับสถานการณ์จริงที่จะเกิดขึ้น

**วิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี**

จากการศึกษาวิธีการจัดการข้อมูลสูญหายประกอบกับการศึกษาจุดเด่นและจุดด้อยของวิธีการต่าง ๆ พบว่า เทคนิควิธีการที่ค่อนข้างมีเหตุผลในการประมาณค่าพารามิเตอร์ได้ถูกต้องแม่นยำก็คือ การทำซ้ำ (Iteration) โดยเฉพาะวิธี EM algorithm จะได้รับการตรวจสอบและยืนยันจากรายงานการวิจัยมากที่สุดว่าเป็นวิธีการที่ดีแต่วิธีการดังกล่าวก็มีจุดบกพร่องอยู่หลายประการ เช่น

1. การแทนค่าสูญหายครั้งแรกในสมการ $\hat{Y} = a + bX$ ซึ่ง $a = \overline{Y} - b\overline{X}$ และ $b = r_{xy}\frac{S_y}{S_x}$ ประกอบไปด้วยค่าสถิติ คือ $\overline{X}, \overline{Y}, r_{xy}, S_X$ และ $S_y$ ได้มาจากกลุ่มตัวอย่างที่มีข้อมูลสมบูรณ์ได้ตัดหน่วยตัวอย่างที่มีข้อมูลสูญหายออกไป ดังนั้น การประมาณค่าพารามิเตอร์ด้วยค่าเหล่านี้จึงทำให้มีอคติ (bias) หมายความว่าค่าที่ได้อาจจะไม่ใช่ค่าของประชากรที่แท้จริง

2. วิธีการจัดการข้อมูลสูญหายแบบ EM algorithm ใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธี Maximum likelihood ซึ่งก็มีข้อจำกัดในการประมาณค่าพารามิเตอร์ดังที่ ราจเมเคอร์ (Raaijmakers, 1999. p. 722) กล่าวไว้ว่า ข้อจำกัดของวิธีการนี้คือใช้เทคนิคการทำซ้ำ (Iteration algorithms) ซึ่งใช้เวลาและเสียค่าใช้จ่ายค่อนข้างสูง และไม่สามารถนำไปรวมไว้กับโปรแกรมคอมพิวเตอร์มาตรฐานทั่ว ๆ ไป และประเด็นที่ถูกวิพากษ์วิจารณ์ก็คือความคลาดเคลื่อนมาตรฐานไม่ตรง (invalid) คือไม่ได้คำนวณอย่างตรงไปตรงมาในการวิเคราะห์ขั้นสุดท้ายดังนั้นจะมีวิธีการอื่น ๆ ที่มีความตรงในการประมาณค่าความคลาดเคลื่อนมาตรฐาน แต่วิธีนี้นักสถิติแนะนำในการใช้แก้ปัญหาข้อมูลสูญหายเพราะใช้วิธีการถดถอยซึ่งน่าจะแทนค่าข้อมูลสูญหายได้ถูกต้องมากยิ่งขึ้น

ดังนั้นผู้วิจัยจึงขอเสนอวิธีการประมาณค่าข้อมูลสูญหายโดยมีขั้นตอนดังต่อไปนี้

1. ประมาณค่าข้อมูลสูญหายโดยวิธีการถดถอยอย่างง่าย (Simple regression) เป็นเทคนิคพื้นฐานของวิธีการอื่น ๆ (Little & Rubin, 1987. p. 61) จากสมการ $\hat{Y} = a + bX$ ใช้ค่าสถิติคือ $a = \overline{Y} - b\overline{X}$ และ $b = r_{xy}\frac{S_y}{S_x}$ เป็นค่าโดยประมาณแต่ถ้าเราสามารถประมาณค่าพารามิเตอร์ได้ ก็จะสามารถทำให้สมการ $\hat{Y} = a + bX$ เป็นสมการที่แท้จริงของประชากร การประมาณค่าข้อมูลสูญหายก็จะได้ค่าที่ถูกต้อง ดังนั้นค่าสถิติ $\overline{X}, \overline{Y}, r_{xy}, S_X$ และ $S_y$ จะได้มาโดยการประมาณค่าแบบทำซ้ำตามทฤษฎีกล่าวว่าค่าคาดหวังของค่าเฉลี่ยจะเท่ากับค่าเฉลี่ยของประชากร $\left(E(\overline{X}) = \mu\right)$ หมายความว่า ถ้ามีค่าเฉลี่ยจำนวนมากแล้วนำมาหาค่าเฉลี่ยจะได้ค่าเฉลี่ยของประชากร ดังนั้น ในขั้นตอนแรกนี้จะทำการ

ตัวอย่างซ้ำจากข้อมูลที่เหลือเมื่อตัดข้อมูลสูญหายออกไปแล้วเป็นจำนวน 1,000 ครั้ง แล้วนำค่าเฉลี่ยทั้ง 1,000 ครั้งมาหาค่าเฉลี่ยก็จะได้ค่าเฉลี่ยของประชากร

การประมาณค่า $S_x$ และ $S_y$ ก็ทำในลักษณะเดียวกันเพราะ $E(S^2)=\sigma^2$

(ประชุม สุวัตถี, 2527. หน้า 24-25)

การประมาณค่าความแปรปรวนดังกล่าวสอดคล้องกับการประมาณค่าความแปรปรวนที่เสนอโดยลิทเทิลและรูบิน (Little & Rubin. 1987. p. 67) ที่กล่าวว่าถ้าให้ $\hat{\theta}_1,...,\hat{\theta}_k$ เป็นตัวแปรสุ่ม ซึ่งไม่มีความสัมพันธ์ต่อกันและมีค่าเฉลี่ยคือ $\mu$ จะได้ว่า $\bar{\theta}=\sum_{j=1}^{k}\frac{\hat{\theta}_j}{k}$ ดังนั้นในการประมาณค่าความแปรปรวนกรณีที่มีข้อมูลสูญหายก็สามารถทำได้โดยการสุ่มข้อมูลที่สมบูรณ์จำนวน 1,000 ครั้ง ในแต่ละครั้งหาความแปรปรวนแล้วนำความแปรปรวนทั้งหมดนั้นมาหาค่าเฉลี่ยก็จะได้ตัวประมาณค่าความแปรปรวนของประชากร

สำหรับการประมาณค่าความสัมพันธ์ของประชากรก็จะทำในลักษณะเดียวกันด้วยเหตุผลต่อไปนี้ (ส่องศรี พิทยารัตน์ และคณะ, 2535. หน้า 76-77) สมการแสดงความแปรปรวนร่วมของกลุ่มตัวอย่าง

$$\begin{aligned}COV(X,Y) &= E(X-\bar{X})(Y-\bar{Y})\\ &= E(XY-\bar{X}Y-\bar{Y}X+\bar{X}\bar{Y})\\ &= E(XY)-\bar{X}E(Y)-\bar{Y}E(X)+\bar{X}\bar{Y}\\ &= E(XY)-\bar{X}\bar{Y}-\bar{Y}\bar{X}+\bar{X}\bar{Y}\\ &= E(XY)-\bar{X}\bar{Y}\end{aligned}$$

ดังนั้นสามารถหาค่าคาดหวังความแปรปรวนร่วมของกลุ่มตัวอย่างได้ดังนี้

$$E(COV(X,Y)) = E(E(XY)-\bar{X}\bar{Y})$$

$$E(COV(X,Y)) = E(E(XY)) - E(\bar{X}\bar{Y})$$

ถ้า a และ b เป็นค่าคงที่ จะได้ E(aX+b)=aE(X)+b ถ้า a=0 แล้ว จะได้ E(b)=b (ส่องศรี พิทยารัตน์, 2535. หน้า 67-68) เนื่องจาก E(XY) เป็นค่าคงที่ตัวหนึ่ง ดังนั้น E(E(XY)) = E(XY) และเนื่องจาก $\bar{X}$ และ $\bar{Y}$ เป็นค่าคงที่จึงน่าจะเป็นอิสระต่อกัน ทวี รื่นจินดา (2525. หน้า 136) กล่าวว่าถ้า X และ Y เป็นอิสระต่อกันแล้ว E(XY)=E(X)E(Y) ดังนั้น $E(\bar{X}\bar{Y})=E(\bar{X})E(\bar{Y})$ แต่ $E(\bar{X})=\mu_x$ และ $E(\bar{Y})=\mu_x$ จะทำให้ได้ว่า

$$E(COV(X,Y)) = E(XY)-\mu_x\mu_y = E(X-\mu_x)(Y-\mu_y)$$

ซึ่งสมการดังกล่าวก็คือความแปรปรวนร่วมของประชากร ($\sigma_{xy}$) นั่นเอง

แสดงว่าค่าคาดหวังความแปรปรวนร่วมของกลุ่มตัวอย่างก็คือความแปรปรวนร่วมของประชากรดังสมการ

$$\sigma_{xy} = E\left(X - \mu_x\right)\left(Y - \mu_y\right)$$

$$\sigma_{xy} = E\left(COV(X, Y)\right)$$ (ประชัย เปี่ยมสมบูรณ์, 2527. หน้า 122)

แต่เนื่องจาก $\rho_{xy} = \sigma_{xy} / \sigma_x \sigma_y$

ดังนั้นเมื่อสามารถประมาณค่าความแปรปรวนร่วมของประชากรได้ก็สามารถประมาณค่าความสัมพันธ์ของประชากรได้เช่นเดียวกัน

ในการประมาณค่าความแปรปรวนร่วมของประชากรกรณีที่มีข้อมูลสูญหาย ผู้วิจัยจะสุ่มข้อมูลที่สมบูรณ์มาจำนวน 1,000 ครั้ง แต่ละครั้งหาความแปรปรวนร่วม แล้วนำความแปรปรวนร่วมทั้งหมดมาหาค่าเฉลี่ยก็จะได้ความแปรปรวนร่วมของประชากร แล้วนำมาหาความสัมพันธ์ของประชากรโดยใช้สูตร $\rho_{xy} = \sigma_{xy} / \sigma_x \sigma_y$ ต่อไป

2. จากการประมาณค่าข้อมูลสูญหายโดยการสร้างสมการทำนาย $\hat{Y} = a + bX$ และใช้การประมาณค่าพารามิเตอร์ในขั้นตอนที่ 1 ก็ทำให้มั่นใจได้ว่า การทำนายข้อมูลสูญหายน่าจะถูกต้องแม่นยำ ลักษณะการสร้างสมการทำนายจะต้องทำให้มีความคลาดเคลื่อนน้อยที่สุดทำให้การทำนายค่า Y (ข้อมูลสูญหาย) ได้อย่างไม่มีอคติค่าที่ได้ใกล้เคียงกับค่าของประชากร ซึ่งความแตกต่างระหว่าง $Y - Y_p$ ($Y_p$ เป็นค่าที่ได้จากการทำนาย) เป็นความคลาดเคลื่อน ($e = Y - Y_p$) ค่าความคลาดเคลื่อนมีทั้งค่าที่เป็นบวกและลบ เส้นถดถอยที่ไม่มีอคติคือเส้นที่ให้ค่าผลรวมของความคลาดเคลื่อนที่เป็นบวกและลบเท่า ๆ กัน แต่เนื่องจาก $\sum(Y - Y_p) = 0$ เพื่อตัดค่าเครื่องหมายลบออกไปจึงยกกำลังสองค่าความคลาดเคลื่อนนั่นคือ $\sum(Y - Y_p)^2 = \sum e^2$ มีค่าน้อยที่สุด จึงจะได้สมการทำนายที่ดีที่สุด แต่ลักษณะการประมาณค่าในขั้นตอนที่ 1 สร้างจากข้อมูลที่สมบูรณ์และข้อมูลสูญหายที่ได้จากการทำนายจึงทำให้ไม่สามารถมั่นใจได้ว่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า (Standard error of the estimate) จะมีค่าน้อยที่สุด ดังนั้น จึงจำเป็นต้องเลือกสมการที่มีความคลาดเคลื่อนน้อยที่สุด (ถ้าสร้างสมการถดถอยจากข้อมูลสมบูรณ์จะมีความคลาดเคลื่อนน้อยที่สุดตามวิธีกำลังสองน้อยที่สุด) เพื่อให้สมการทำนายข้อมูลสูญหายได้อย่างถูกต้องแม่นยำ

ดังนั้น เมื่อแทนค่าข้อมูลสูญหายในขั้นตอนที่ 1 แล้ว ผู้วิจัยจึงสร้างสมการการทำนายข้อมูลสูญหายจากข้อมูลทั้งหมด ทั้งข้อมูลสมบูรณ์และข้อมูลสูญหายที่ได้จากการทำนายโดยใช้วิธีการถดถอยอย่างง่าย (Simple regression) จำนวน 1,000 ครั้ง แล้วเลือกสมการที่มีความคลาดเคลื่อนมาตรฐานในการประมาณค่าน้อยที่สุด เป็นสมการทำนายข้อมูลสูญหายในการศึกษาวิจัยครั้งนี้

จากที่กล่าวมาจะเห็นได้ว่าในขั้นตอนที่ 1 เป็นการประมาณค่าพารามิเตอร์ (Estimated Parameter) และในขั้นตอนที่ 2 เป็นการคัดเลือกสมการที่มีความคลาดเคลื่อนมาตรฐานในการประมาณค่าน้อยที่สุด (Smallest Standard Error) ดังนั้น ผู้วิจัยจึงตั้งชื่อวิธีการจัดการข้อมูลสูญหายแบบนี้ว่าวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี

**วัตถุประสงค์ของการวิจัย**

การวิจัยครั้งนี้มีวัตถุประสงค์ของการวิจัยดังต่อไปนี้

1. เพื่อพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี

2. เพื่อตรวจสอบความแม่นยำและอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน โดยพิจารณาเปรียบเทียบดังนี้

2.1 เพื่อเปรียบเทียบความแม่นยำของค่าเฉลี่ยเลขคณิตที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2.2 เพื่อเปรียบเทียบความแม่นยำของความแปรปรวนที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2.3 เพื่อเปรียบเทียบความแม่นยำของค่าสัมประสิทธิ์สหสัมพันธ์ที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2.4 เพื่อเปรียบเทียบอำนาจการทดสอบของข้อมูลที่ได้จากวิธีการจัดการข้อมูลสูญหาย วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

3. เพื่อศึกษาปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ของตัวแปร ที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์

**สมมุติฐานการวิจัย**

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า การศึกษาวิธีการจัดการข้อมูลสูญหายจะมีตัวแปรที่เข้ามาเกี่ยวข้อง คือ วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหาย ดังนั้น ผู้วิจัยจึงตั้งสมมุติฐานการวิจัยดังนี้

1. วิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี น่าจะมีความแม่นยำและอำนาจการทดสอบดีกว่าวิธีอื่นเมื่อใช้วิธีการสุ่มตัวอย่าง ความสัมพันธ์ของตัวแปร และจำนวนข้อมูลสูญหายที่แตกต่างกัน

2. ความแม่นยำน่าจะขึ้นอยู่กับปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ของตัวแปรที่แตกต่างกัน

**ข้อมูล**

ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากการจำลองสถานการณ์โดยใช้เทคนิคมอนติ คาร์โล ซิมูเลชั่น (Monte Carlo Simulation Technique) เป็นข้อมูลผลสัมฤทธิ์ทางการเรียนประกอบด้วยเกรดเฉลี่ยเป็นตัวแปรเกณฑ์ (Y) มีค่าเฉลี่ยเท่ากับ 2.00 และความแปรปรวนเท่ากับ 0.15 และตัวแปรทำนาย คือคะแนนสอบของนักเรียน (X) มีค่าเท่ากับ 50 และความแปรปรวนเท่ากับ 100

**ขนาดของกลุ่มตัวอย่าง**

การกำหนดขนาดกลุ่มตัวอย่างในการวิจัยเชิงสำรวจจะต้องให้มีความเหมาะสมที่ผู้วิจัยสามารถดำเนินการวิจัยได้ซึ่งไม่ควรมากหรือน้อยเกินไป ถ้าน้อยเกินไปการสรุปอ้างอิงไปยังประชากรก็ไม่ถูกต้อง แต่ถ้ามากเกินไปผู้วิจัยก็ไม่สามารถทำได้เพราะต้องลงทุนสูง ดังนั้นเพื่อให้สามารถนำผลการวิจัยไปใช้ได้สอดคล้องกับสภาพความเป็นจริง ผู้วิจัยจึงกำหนดขนาดกลุ่มตัวอย่างเท่ากับ 350 คน ซึ่งสอดคล้องกับงานวิจัยเชิงสำรวจที่ศึกษาข้อมูลสูญหายในต่างประเทศและสอดคล้องกับการกำหนดขนาดกลุ่มตัวอย่างโดยใช้ตารางของยามาเน่และเครจซี่และมอร์แกน

การกำหนดขนาดกลุ่มตัวอย่างย่อยในแต่ละระดับชั้นหรือในแต่ละกลุ่มใช้สูตรของนีย์แมนซึ่งเป็นวิธีการกำหนดขนาดกลุ่มตัวอย่างที่ถือว่าค่าใช้จ่ายต่อหน่วยในการสำรวจแต่ละระดับชั้นมีค่าใกล้เคียงกันหรือไม่แตกต่างกันอย่างเห็นได้ชัดเจนหรือเท่ากัน (สุกัญญรัตน์ คงงาม, 2539. หน้า 22-23) และจากการศึกษาของ ดวงใจ ปวีณอภิชาต (2535) พบว่า การกำหนดขนาดกลุ่มตัวอย่างย่อยแบบนีย์แมนเป็นวิธีการกำหนดขนาดกลุ่มตัวอย่างที่มีประสิทธิภาพมากที่สุดในการประมาณค่าเฉลี่ยเลขคณิตและความแปรปรวนของคะแนนผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ของประชากรการกำหนดขนาดตัวอย่างของแต่ละระดับชั้น มีสูตรดังนี้

$$n_h = \frac{nN_hS_h}{\sum_{h=1}^{L} N_hS_h} \qquad \text{โดยที่ } h=1,2,3,...,L$$

เมื่อ $n_h$ คือ ขนาดของกลุ่มตัวอย่างในระดับชั้นที่ h

$S_h$ คือ ส่วนเบี่ยงเบนมาตรฐานของประชากรย่อยในระดับชั้นที่ h

L คือ จำนวนระดับชั้น

n คือ ขนาดของกลุ่มตัวอย่าง

วิธีดำเนินการ

ผู้วิจัยจำลองสถานการณ์โดยใช้เทคนิคมอนติคาร์โล ซิมูเลชั่น (Monte Carlo Simulation Technique) ด้วยเครื่องคอมพิวเตอร์จุลภาค เขียนโปรแกรมด้วยภาษาฟอกโปร (FOXPRO) ได้ข้อมูลผลสัมฤทธิ์ทางการเรียน ประกอบด้วยเกรดเฉลี่ยเป็นตัวแปรเกณฑ์ (Y) และตัวทำนายคือคะแนนสอบของนักเรียน (X) และผู้วิจัยได้สร้างประชากรขึ้นมาอีก 2 กลุ่ม เพื่อใช้ในการศึกษาค่าอำนาจการทดสอบโดยให้มีค่าเฉลี่ยของเกรดเฉลี่ยมากกว่าค่าเฉลี่ยของเกรดเฉลี่ยจากประชากรในการจำลองสถานการณ์ครั้งแรกเท่ากับ $1\sigma$ และน้อยกว่าค่าเฉลี่ยของเกรดเฉลี่ยจากประชากรในการจำลองครั้งแรกเท่ากับ $-1\sigma$ แล้วจึงนำข้อมูลจากประชากรทั้งหมดมากำหนดว่าแต่ละคนเป็นกลุ่มใดและระดับชั้นใด

ดำเนินการวิเคราะห์ข้อมูลจากประชากรกลุ่มแรกที่มีความสัมพันธ์ต่ำ (r=.30) โดยหาค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ แล้วจึงสุ่มตัวอย่างโดยใช้คอมพิวเตอร์สุ่มจากประชากรกลุ่มนี้ด้วยวิธีการสุ่มตัวอย่างแบบแบ่งชั้น จำนวน 350 คน สร้างข้อมูลสูญหายแบบสุ่ม (randomly missing data) โดยการเขียนคำสั่งคอมพิวเตอร์ให้ลบข้อมูลเกรดเฉลี่ยออกไปเท่ากับ 5% ดำเนินการจัดกระทำข้อมูลสูญหายโดยการตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ คำนวณหาค่าเฉลี่ยความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ และคำนวณความแม่นยำโดยพิจารณาจากความแตกต่างของค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ ที่ได้จากการสุ่มตัวอย่างแบบแบ่งชั้นและตัดข้อมูลสูญหายออกแบบลิสท์ไวส์กับค่าที่คำนวณจากประชากร หลังจากนั้นจึงคำนวณอำนาจการทดสอบจากการทดสอบความสัมพันธ์

เปลี่ยนกลุ่มประชากรเป็นกลุ่มที่สองซึ่งมีค่าเฉลี่ยของเกรดเฉลี่ยมากกว่าประชากรกลุ่มแรกเท่ากับ $1\sigma$ สุ่มตัวอย่างโดยใช้คอมพิวเตอร์สุ่มจากประชากรกลุ่มนี้ด้วยวิธีการสุ่มแบบแบ่งชั้น จำนวน 350 คน สร้างข้อมูลสูญหายแบบสุ่ม (randomly missing data) จำนวน 5% จัดกระทำข้อมูลสูญหายโดยการตัดออกแบบลิสท์ไวส์ แล้วคำนวณอำนาจการทดสอบจากการทดสอบที (t-test) โดยใช้ข้อมูลจากการสุ่มประชากรกลุ่มแรกและประชากรกลุ่มที่สอง เปลี่ยนกลุ่มประชากรเป็นกลุ่มที่สามซึ่งมีค่าเฉลี่ยของเกรดเฉลี่ยน้อยกว่าประชากรกลุ่มแรกเท่ากับ $-1\sigma$ สุ่มตัวอย่างโดยใช้การสุ่มแบบแบ่งชั้น จำนวน 350 คน สร้างข้อมูลสูญหายแบบสุ่ม จำนวน 5% แล้วจึงตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ คำนวณอำนาจการทดสอบจากการทดสอบที (t-test) โดยใช้ประชากรกลุ่มที่สองกับกลุ่มที่สามมาเปรียบเทียบกัน และคำนวณอำนาจการทดสอบจากการทดสอบเอฟ (F-test) หรือการวิเคราะห์ความแปรปรวน โดยใช้ข้อมูลจากการสุ่มประชากรกลุ่มแรก กลุ่มที่สองและกลุ่มที่สาม มาเปรียบเทียบกัน ดำเนินการทำซ้ำแบบเดิมดังที่กล่าวมาจำนวน 1,000 ครั้ง แล้วจึงคำนวณหาค่าเฉลี่ยกำลังสองความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ คำนวณอำนาจการทดสอบโดยพิจารณาสัดส่วนจากจำนวนครั้งของการปฏิเสธสมมุติฐานศูนย์ของการทดสอบความสัมพันธ์ การทดสอบที (t-test) และการทดสอบเอฟ (F-test) กับจำนวนครั้งของการทดสอบทั้งหมด

ดำเนินการวิเคราะห์ข้อมูลในลักษณะเดียวกัน โดยกำหนดความสัมพันธ์ของตัวแปรเท่ากับ .50 และ .70 ใช้วิธีการสุ่มตัวอย่างแบบกลุ่มและแบบหลายขั้นตอน จำนวนข้อมูลสูญหายเท่ากับ 10% 20% และ 30% วิธีการจัดการข้อมูลสูญหายโดยการแทนค่าแบบอีเอ็มและแบบอีพีเอสเอสอี ได้สถานการณ์จำลอง 108 เงื่อนไขการทดลองในแต่ละสถานการณ์จะมีข้อมูล 1,000 กรณี

**การวิเคราะห์ข้อมูล**

การวิจัยนี้ใช้แบบแผนการวิจัยเชิงทดลอง 3 X 3 X 3 X 4 ซึ่งมีตัวแปรอิสระ 4 ตัวแปร คือ 1). วิธีการสุ่ม ตัวอย่าง 3 วิธี 2). วิธีการจัดการข้อมูลสูญหาย 3 วิธี 3). ความสัมพันธ์ของตัวแปร 3 ระดับ และ 4). จำนวนข้อมูลสูญหาย 4 ระดับ ตัวแปรตาม คือ ความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวนและสัมประสิทธิ์สหสัมพันธ์ และอำนาจการทดสอบ วิเคราะห์ข้อมูลโดยการนำค่าความแม่นยำที่ได้ในแต่ละการทดลองมาเปรียบเทียบกันโดยถือว่าค่าที่ได้จากการทำซ้ำ 1,000 ครั้ง เป็นค่าพารามิเตอร์มีความคงที่ไม่เปลี่ยนแปลง ส่วนการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหายและความสัมพันธ์ของตัวแปร ที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ ใช้การวิเคราะห์ ความแปรปรวน 4 ทาง (Four-Way ANOVA) เมื่อเกิดปฏิสัมพันธ์ระหว่างตัวแปร ผู้วิจัยได้ค้นหาคำตอบเกี่ยวกับการเกิดปฏิสัมพันธ์โดยการวิเคราะห์ Simple Interaction Effect , Simple Main Effect และ Simple Simple Effect

**ผลการวิจัย**

ผลการวิจัยสรุปได้ดังต่อไปนี้

1. วิธีการจัดการข้อมูลสูญหายโดยการแทนค่าแบบอีพีเอสเอสอีได้ค่าความแม่นยำของค่าเฉลี่ยเลขคณิตไม่แตกต่างจากวิธีอีเอ็ม อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเป็นค่าที่สูงที่สุด เมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่มและแบบหลายขั้นตอน จำนวนข้อมูลสูญหายอยู่ในระดับสูงที่สุดคือเท่ากับ 30% ความสัมพันธ์ระหว่างตัวแปรอยู่ในระดับสูง (r=.70)

2. วิธีการจัดการข้อมูลสูญหายโดยการตัดออกแบบลิสท์ไวส์ได้ค่าความแม่นยำของความแปรปรวนแตกต่างจากวิธีอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเป็นค่าที่สูงที่สุดเมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่มและแบบหลายขั้นตอน จำนวนข้อมูลสูญหาย 10% 20% และ 30% ความสัมพันธ์ระหว่างตัวแปรอยู่ในระดับต่ำ (r=.30) ปานกลาง (r=.50) และสูง (r=.70)

3. วิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์ ได้ค่าความแม่นยำของสัมประสิทธิ์สหสัมพันธ์แตกต่างจากวิธีการจัดการข้อมูลสูญหายแบบอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเป็นค่าที่

สูงที่สุด เมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่ม และแบบหลายขั้นตอน จำนวนข้อมูลสูญหายอยู่ในระดับสูงที่สุดคือเท่ากับ 30% ความสัมพันธ์ระหว่างตัวแปรอยู่ในระดับสูง (r=.70)

4. วิธีการจัดการข้อมูลสูญหายโดยการแทนค่าแบบอีพีเอสเอสอี ได้ค่าอำนาจการทดสอบ เมื่อใช้การทดสอบความสัมพันธ์สูงกว่าวิธีการจัดการข้อมูลสูญหายโดยการตัดออกแบบลิสท์ไวส์และการแทนค่าแบบอีเอ็ม เมื่อใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้น จำนวนข้อมูลสูญหายทุกระดับ (5%,10%,20% และ 30%) ความสัมพันธ์ระหว่างตัวแปรอยู่ในระดับต่ำ (r=.30)

5. ปฏิสัมพันธ์สี่ทางระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหายและความสัมพันธ์ระหว่างตัวแปรที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ ไม่มีนัยสำคัญทางสถิติที่ระดับ .01

6. ปฏิสัมพันธ์สามทางระหว่าง 1). วิธีการสุ่มตัวอย่างกับวิธีการจัดการข้อมูลสูญหายและจำนวนข้อมูลสูญหาย 2). วิธีการสุ่มตัวอย่างกับวิธีการจัดการข้อมูลสูญหายและความสัมพันธ์ระหว่างตัวแปร 3). วิธีการจัดการข้อมูลสูญหายกับจำนวนข้อมูลสูญหายและความสัมพันธ์ระหว่างตัวแปร ที่มีต่อความแม่นยำของความแปรปรวนมีนัยสำคัญทางสถิติที่ระดับ .01 และปฏิสัมพันธ์สามทางระหว่างระหว่างวิธีการสุ่มตัวอย่างกับวิธีการจัดการข้อมูลสูญหายและจำนวนข้อมูลสูญหาย ที่มีต่อความแม่นยำของสัมประสิทธิ์สหสัมพันธ์มีนัยสำคัญทางสถิติที่ระดับ .01

## บรรณานุกรม


ดวงใจ ปวีณอภิชาต. (2535). การเปรียบเทียบค่าประมาณพารามิเตอร์ของแบบแผนการสุ่มแบบแบ่งชั้นที่มีตัวแปรจำแนกชั้นภูมิ และวิธีการกำหนดขนาดของกลุ่มตัวอย่างย่อยที่แตกต่างกัน : กรณีศึกษาผลสัมฤทธิ์ทางการเรียน. วิทยานิพนธ์ ค.ม., จุฬาลงกรณ์มหาวิทยาลัย.

ประชุม สุวัตถี. (2527). ทฤษฎีการอนุมานเชิงสถิติ. กรุงเทพมหานคร: สำนักพิมพ์โอเดียนสโตร์.

ปุระชัย เปี่ยมสมบูรณ์. (2527). การวิเคราะห์เส้นโยงทางสังคมและพฤติกรรมศาสตร์. กรุงเทพมหานคร: สำนักพิมพ์โอเดียนสโตร์.

สุกัญญรัตน์ คงงาม. (2539). การเปรียบเทียบคุณสมบัติของตัวประมาณค่าพารามิเตอร์ที่ได้จากกลุ่มตัวอย่างสุ่มแบบหลายขั้นตอน ระหว่างวิธีสุ่มแบบง่ายกับแบบมีระบบ. วิทยานิพนธ์ ค.ม., จุฬาลงกรณ์มหาวิทยาลัย.

ส่องศรี พิทยารัตน์ และคณะ. (2535). หลักสถิติ. กรุงเทพมหานคร: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

Little, R. J. A. & Rubin, D. B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Raaijmakers, Q. A. W. (1999). Effectiveness of different missing data treatments in surveys with likert-type data : Introducing the relative mean substitution approach. Educational and Psychological Measurement, 59(5), 725-748.

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