บทที่ 2

เอกสาร และผลงานวิจัยที่เกี่ยวข้อง

ขนาดของกลุ่มตัวอย่างและการสูญหายของข้อมูล

ในการวิจัยเชิงทดลองโดยเฉพาะทางจิตวิทยานั้นส่วนมากมักจะเผชิญกับปัญหาในด้านข้อจำกัดเกี่ยวกับขนาดของกลุ่มตัวอย่างที่ใช้ในการทดลอง คือมักเป็นการทดลองกับกลุ่มตัวอย่างขนาดเล็กหรือมีจำนวนไม่มากนัก เนื่องจากสาเหตุหลาย ๆ ประการ (Sims 1973: 2-3) คือ

1. เป็นการศึกษารายกรณีหรือกลุ่มย่อย ๆ ที่ผู้วิจัยจะทำการศึกษาสภาพการณ์ต่าง ๆ อย่างเข้มข้นและลึกซึ้ง (Intensive Study) เพื่อให้ตัวแปรทดลองส่งผลต่อตัวแปรตามมากที่สุด และสามารถควบคุมตัวแปรแทรกซ้อนที่อาจจะส่งผลต่อตัวแปรตามด้วย

2. ลักษณะตามธรรมชาติของข้อมูลมีจำนวนน้อย เพราะการศึกษาลักษณะทางจิตวิทยาบางสภาพการณ์ ผู้วิจัยไม่สามารถที่จะหาจำนวนของตัวอย่างหรือข้อมูลที่มีสภาพตามที่ต้องการได้มากเพียงพอ

3. ประหยัดในด้านค่าใช้จ่ายจากเครื่องมือและอุปกรณ์ที่ใช้ในการทดลอง เพราะบางงานวิจัยต้องศึกษากันในระยะยาว (Longitudinal Study) เพื่อให้สามารถสรุปผลการวิจัยที่แน่ชัดและแม่นยำว่าผลที่เกิดขึ้นนั้น เนื่องจากตัวแปรทดลองอย่างแน่นอน

จากเหตุผลดังกล่าว การวิจัยเชิงทดลองจึงใช้ข้อมูลจำนวนน้อยเป็นกลุ่มขนาดเล็ก ดังนั้น เมื่อมีข้อมูลบางส่วนสูญหายไปจากการทดลอง จึงเกิดปัญหาในการสรุปผลเป็นอย่างมาก เพราะว่าจะได้ข้อมูลจากกลุ่มตัวอย่างลดลงไป ซึ่งมีผลต่อค่าความแปรปรวนและความถูกต้องแม่นยำในการวิเคราะห์ข้อมูล (Morrison 1976: 120) ซึ่งการสูญหายของข้อมูลเกิดจากการที่ผู้วิจัยไม่สามารถเก็บรวบรวมข้อมูลเกี่ยวกับตัวแปรที่ศึกษาได้ครบถ้วนเมื่อเสร็จสิ้นงานวิจัย โดยเกิดขึ้นได้ในหลายกรณีต่อไปนี้เช่น พลวิจัยเกิดการเจ็บป่วยหรือตายในระหว่างทำการทดลอง หรือเกิดจากการอพยพย้ายสถานที่อยู่ หรือเกิดจากการไม่ตอบสนองของพลวิจัยต่อการวัด หรือการเก็บรวบรวมข้อมูล หรือเกิดจากการที่ผู้วิจัยไม่ได้บันทึกข้อมูลในวาระนั้น

เวลช์ แฟรงค์ และคอสเทลโล (Welch, Frank and Costello 1983: 177-180) กล่าวว่ากรณีข้อมูลสูญหายเป็นปัญหาที่สำคัญมากอย่างหนึ่งในงานวิจัยทางด้านจิตบำบัด (Psychiatic Research) มักเกิดขึ้นเสมอ ๆ ในกลุ่มตัวอย่างที่ทำการศึกษา และที่สำคัญประการหนึ่งก็คือข้อมูลที่สูญหายนั้นไม่เป็นลักษณะแบบสุ่ม จะเฉพาะเจาะจงกับคนไข้ที่มีปัญหามาก ๆ เท่านั้น ทำให้การวิเคราะห์ทางสถิติในการสรุปผลและการอภิปรายเกิดลำเอียงอย่างเห็นได้ชัด

โคเฮน และโคเฮน (Cohen and Cohen 1983: 278) กล่าวว่าถ้าข้อมูลที่สูญหายมีลักษณะแบบสุ่ม การทดสอบความสัมพันธ์ของผลลัพธ์ที่ได้จะไม่ลำเอียงไปจากสมมติฐานขณะเมื่อมีข้อมูลครบถ้วน แต่จะมีผลทำให้เกิดความคลาดเคลื่อนของชั้นแห่งความเป็นอิสระเป็นผลให้อำนาจการทดสอบลดต่ำลง และยังมีผลให้ค่าความคลาดเคลื่อนมาตรฐานใหญ่ขึ้น คือไปลดค่าความแม่นตรง (precision)

ซาร์โควิช (Zarkovich 1966: 148-149) ได้ทำการวิจัยเชิงสำรวจ ซึ่งในขั้นการเก็บรวบรวมข้อมูลเขาได้พบว่ามีข้อมูลสูญหายเป็นจำนวนมาก และเขาได้ใช้วิธีประมาณค่าแทนข้อมูลที่สูญหาย พบว่าการประมาณค่าแทนข้อมูลที่สูญหาย จะมีผลต่อค่าของความแปรปรวนมากกว่าการตัดข้อมูลที่สูญหายออกจากการวิเคราะห์ข้อมูล คือทำให้ได้ค่าความแปรปรวนที่ใกล้เคียงกับความเป็นจริงตรงกับข้อมูลที่เขาเก็บรวบรวมเพิ่มจากครั้งก่อน

แบบบี (Babbie 1986: 372-373) ได้ให้ข้อเสนอแนะแนวทางปฏิบัติเกี่ยวกับข้อมูลที่สูญหายไว้หลายวิธีคือ ถ้าผู้วิจัยมีข้อมูลที่สัมพันธ์กับที่สูญหายสัก 2-3 ข้อมูล และข้อมูลที่ใช้ในการวิเคราะห์มีจำนวนมากเพียงพอ ควรแยกข้อมูลที่สัมพันธ์กับที่สูญหายออกจากการวิเคราะห์ข้อมูลเบื้องต้นด้วย หรือผู้วิจัยควรวางแนวทางปฏิบัติเตรียมไว้สำหรับกรณีมีข้อมูลสูญหายเกิดขึ้น และข้อสำคัญควรระมัดระวังในการวิเคราะห์และตีความหมายจากข้อมูลที่มีการสูญหาย แบบบีได้ศึกษาต่อมา พบว่าถ้าจะอนุรักษ์ผลการวิเคราะห์ข้อมูล (Conservative) การประมาณค่าแทนข้อมูลที่สูญหายจะให้ผลการวิเคราะห์ดีกว่าการตัดข้อมูลในส่วนที่สมบูรณ์ของกลุ่มอื่นออก เพื่อให้เหลือข้อมูลในแต่ละกลุ่มการทดลองเท่ากัน หรือวิเคราะห์ข้อมูลไปโดยไม่คำนึงถึงข้อมูลส่วนที่สูญหาย

คอคแครน และคอกซ์ (Cochran and Cox 1957: 80-82) ได้แนะนำให้ผู้วิจัยเลือกใช้วิธีประมาณค่าแทนข้อมูลที่สูญหาย ซึ่งสอดคล้องกับกลีสันและสเตลิน (Gleason and Staelin 1975: 230-231) มอร์ริสัน (Morrison 1976: 120) มาร์ชูโล (Marascuilo 1983: 65-66) เคปเพล (Keppel 1982: 100-101) ที่เสนอแนะให้ผู้วิจัยเลือกใช้การประมาณค่าแทนข้อมูลสูญหายเช่นกัน

ยัง (Young 1981: 367) กล่าวว่าเพื่อให้การวิเคราะห์ข้อมูลมีคุณภาพสูงที่สุด ผู้วิจัยจำเป็นต้องมีความคิดรวมยอด และศึกษาค้นคว้าเกี่ยวกับสถานภาพทั่ว ๆ ไปของตัวข้อมูลที่สูญหายให้มากที่สุด เพื่อสามารถจะอธิบายเกี่ยวกับข้อมูลที่สูญหายได้อย่างเต็มที่ ก่อนที่จะเลือกใช้วิธีประมาณค่าวิธีหนึ่งวิธีใด

<u>วิธีประมาณค่าข้อมูลที่สูญหาย 3 วิธี</u>

1. วิธีประมาณค่าข้อมูลที่สูญหายโดยใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง (Estimated by Sampling mean)

ถ้าผู้วิจัยกำหนดขนาดกลุ่มตัวอย่างในการวิจัยไว้เป็น n แต่เมื่อสิ้นสุดการวิจัยปรากฏว่ามีข้อมูลสูญหายไปและสามารถเก็บรวบรวมข้อมูลได้เพียง n_x หน่วย โดยที่ n_x มีค่าน้อยกว่า n จากสถานการณ์นี้สามารถประมาณค่าแทนข้อมูลที่สูญหายไป $(n - n_x)$ หน่วยได้ด้วยค่า $\bar{X}_{n_x}$ ซึ่งเป็นวิธีประมาณค่าโดยใช้ค่าเฉลี่ยจากกลุ่มตัวอย่างที่เหลือ นักสถิติส่วนใหญ่จะรู้รู้จักกันดีในชื่อของ The Zero Order method กล่าวได้ว่าเป็นเทคนิควิธีที่ง่ายไม่มีขั้นตอนยุ่งยากซับซ้อนในการคำนวณ ในปี ค.ศ. 1959 เดียร์ (Dear, R.E. 1959, Cited by Afifi, A.A. and Elashoff, R.M. 1966: 598-599) เป็นผู้เสนอการอภิปรายการคำนวณสำหรับวิธีนี้ โดยคำนวณหาค่าเฉลี่ยจากข้อมูลที่มีอยู่ทั้งหมดแล้วนำค่าที่ได้แทนค่าของข้อมูลที่สูญหาย ซึ่งเขียนอยู่ในรูปสมการได้ดังนี้

$$\bar{X}_{n_x} = \frac{\sum_{i=1}^{n_x} X_i}{n_x}$$

เมื่อ $\bar{X}_{n_x}$ คือ ค่าเฉลี่ยจากข้อมูลที่ยังคงเหลืออยู่

n_x คือ จำนวนหน่วยตัวอย่างที่มีค่าของคะแนน X ทั้งหมด

2. วิธีประมาณค่าข้อมูลที่สูญหายโดยใช้สมการถดถอย (Estimated by Regression equation)

ในปี ค.ศ. 1966 อะฟิฟี และอิลาสฮอฟฟ์ (Afifi and Elashoff 1966: 599-600) ได้เสนอวิธีประมาณค่าแทนข้อมูลที่สูญหายด้วยสมการถดถอย โดยใช้ชื่อว่า The First Order Regression เป็นวิธีประมาณค่าที่ตัวแปรต้นและตัวแปรตามหรือตัวแปรควบคุม (Covariate Variable) มีความสัมพันธ์กันในลักษณะ เชิงเส้น เท่านั้น ในกรณีที่ตัวแปรตามมีข้อมูลสูญหายไปจะสามารถประมาณค่าได้จากตัวควบคุมที่มีข้อมูล เหลืออยู่

ทาแบชนิค และฟีเดล (Tabachnick and Fidell 1983: 72) ได้เสนอแนะว่า การใช้สมการถดถอยสำหรับทำนายค่าที่ขาดหายไปควรจะมีตัวทำนายที่ดีที่มีความสัมพันธ์สูง ๆ กับตัวแปรที่ขาดหายไป โดยใช้ค่าประมาณที่ตกอยู่ในช่วงของพิสัยของค่ารวม ๆ ระหว่างค่าที่ใช้สร้างสมการถดถอย เท่านั้น และควรใช้ข้อมูลจากกลุ่มตัวอย่างที่มีสภาพคล้ายคลึงกันและมีการสูญหายของข้อมูลไม่มากนัก

ในการทดลองหนึ่งที่ขนาดกลุ่มตัวอย่าง เท่ากับ 10 ผู้วิจัย เก็บรวบรวมข้อมูลของตัวแปรตาม (Y) ได้บางส่วน ซึ่งเขามีข้อมูลของตัวแปรควบคุม (X) ที่มีความสัมพันธ์ เชิงเส้นกับตัวแปรตามดังนี้

ตัวแปร	1	2	3	4	5	6	7	8	9	10
ตัวแปรควบคุม	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
ตัวแปรตาม	Y1	Y2	Y3	—	Y5	Y6	—	Y8	Y9	Y10

ผู้วิจัยสามารถประมาณค่า Y4 และ Y7 แทนข้อมูลที่สูญหายด้วยสมการถดถอย โดยมีวิธีการคำนวณดังนี้

$$Y_i = \alpha + \beta X_i \qquad \text{.....(1)}$$

โดยใช้ เทคนิค Least Square Method คำนวณหาค่า α และ β ได้ดังนี้

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} Y_i - \beta \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\alpha = \bar{Y} - \beta\bar{X}$$

และ

$$\beta = \frac{\sum_{i=1}^{N} Y_i X_i - \bar{Y}\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} X_i^2 - \bar{X}\sum_{i=1}^{N} X_i}$$

$$\beta = \frac{\sum_{i=1}^{N} Y_i X_i - N\bar{Y}\bar{X}}{\sum_{i=1}^{N} X_i^2 - N\bar{X}^2}$$

(Mararcuilo 1971: 480-481)

ต่อจากนั้นนำค่า α , β และ X4 แทนลงในสมการ (1) เพื่อคำนวณหาค่า $\hat{Y}4$ แล้วนำไปแทนตัวที่ขาดหายไป ซึ่งค่าที่เหลือคือ $\hat{Y}7$ ก็ทำนองเดียวกัน

3. วิธีประมาณค่าข้อมูลที่สูญหายโดยใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยจากกลุ่มตัวอย่างและสมการถดถอย (Estimated by mean of Sampling mean and Regression equation)

ในปี ค.ศ. 1967 อะฟีฟี และอิลาสฮอฟฟ์ เป็นผู้คิดค้นเสนอวิธีประมาณค่านี้ไว้ (Afifi and Elashoff 1967: 16) โดยใช้แนวความคิดที่ว่า วิธีประมาณค่าข้อมูล โดยใช้สมการถดถอยช่วยทำนาย เป็นวิธีที่มีลักษณะ เฉพาะที่ดี คือใช้ค่าสัมประสิทธิ์สหสัมพันธ์ในการทำนายข้อมูลได้อย่างดี เมื่อตัวแปรต้นและตัวแปรตามมีความสัมพันธ์กัน และสำหรับวิธีประมาณค่าข้อมูลโดยใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง เป็นวิธีที่มีคุณสมบัติลดค่าความแปรปรวน ดังนั้น อะฟีฟี และอิลาสฮอฟฟ์ จึงได้รวมวิธีประมาณค่าทั้งสองเข้าด้วยกัน โดยใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยจากกลุ่มตัวอย่างและค่าจากสมการถดถอย ซึ่งแทนค่าที่สูญหายด้วยค่าเฉลี่ยก่อนคำนวณค่าจากสมการถดถอย วิธีประมาณค่านี้มีลำดับขั้นในการคำนวณดังนี้

1. คำนวณหาค่าเฉลี่ยจากกลุ่มตัวอย่าง จะได้ค่า

2. แทนข้อมูล X_i ที่สูญหายไปด้วย $\bar{X}_{n_x}$ แล้วนำข้อมูลที่ได้ทั้งหมดคำนวณการถดถอยของ X ในเทอม Y โดยวิธีกำลังสองน้อยที่สุด จะได้สมการถดถอย

$$\text{คือ} \quad X_i^{LS} = d_o^{LS} + d_1^{LS} Y_i$$

3. คำนวณหาค่า X_i ที่สูญหายโดยใช้สมการจากข้อ 2 จะได้ค่า $\hat{X}_i^{LS}$

4. หาค่า $(\bar{X}_{n_x} + \hat{X}_i^{LS})/2$ ซึ่งเป็นค่าประมาณที่ได้จากวิธีประมาณค่านี้

<u>คุณสมบัติของวิธีประมาณค่าที่ดี</u>

ในการประมาณค่าพารามิเตอร์หรือลักษณะต่าง ๆ ของประชากรที่ใช้ในการวิเคราะห์ข้อมูล วิธีที่ใช้ในการประมาณค่ามี 2 แบบ คือ การประมาณค่าแบบจุดและการประมาณค่าแบบช่วง (point and interval estimates) สำหรับวิธีการประมาณค่าแบบจุดเป็นการประมาณค่าพารามิเตอร์ของประชากรที่สนใจศึกษาด้วยค่าเพียงค่าเดียวเท่านั้น เช่น ใช้ $\bar{X}$ ประมาณค่า μ หรือใช้ s^2 ประมาณค่า σ^2 เป็นต้น ส่วนวิธีการประมาณค่าแบบช่วงเป็นการประมาณค่าพารามิเตอร์ของประชากรที่สนใจศึกษาด้วยช่วงค่าช่วงหนึ่ง ซึ่งมีคุณสมบัติว่าค่าของประชากรที่แท้จริงจะตกอยู่ในช่วงค่าที่ประมาณนี้ ด้วยความเชื่อมั่นระดับหนึ่ง โดยจะต้องอาศัยการประมาณค่าแบบจุด และการแจกแจงความน่าจะเป็นของตัวประมาณเป็นพื้นฐานในการคำนวณ การประมาณค่าทั้งสองแบบจะเหมาะสมกับกรณีการใช้งานที่ต่างกัน กรณีที่มีตัวประมาณค่า (Estimators) อยู่หลายวิธีที่สามารถนำมาใช้ในการประมาณค่าสิ่งใดสิ่งหนึ่งได้ จึงมีการกำหนดคุณสมบัติของวิธีประมาณค่าที่ดีควรมีคุณสมบัติครบ 4 ประการดังต่อไปนี้ (Hays 1963: 196-201 Yamane 1967: 239-245, and Wilks 1962: 256-261)

1. ความไม่เอนเอียง (Unbiasness) หมายถึง ถ้า $\hat{\theta}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของพารามิเตอร์ θ แล้ว

จะได้ว่า $E(\hat{\theta}) = \theta$

นั่นคือ ค่าคาดหวัง (Expected Value) ของตัวประมาณค่า θ มีค่าเท่ากับค่าของพารามิเตอร์ตัวที่ต้องการประมาณหรือกล่าวอีกนัยหนึ่งได้ว่า ตัวประมาณค่านั้นไม่เอนเอียงถ้า Expectation ของการแจกแจงของ $\hat{\theta}$ มีค่าเท่ากับพารามิเตอร์ θ และตัวประมาณค่าที่ไม่

เอนเอียงนั้นมีคุณสมบัติที่ดีอยู่ว่า ถ้ามีชุดของค่าประมาณที่ไม่เอนเอียงที่เป็นอิสระต่อกันอยู่แล้ว ค่าเฉลี่ยของค่าเหล่านั้นย่อมไม่เอนเอียงด้วย และในทางตรงข้ามกันค่าเฉลี่ยของค่าประมาณที่เอนเอียงย่อมเอนเอียงด้วยไม่ว่าจะเฉลี่ยมาจากกี่ค่าก็ตาม

2. ความสอดคล้อง (Consistency) หมายถึง ถ้าประมาณค่าพารามิเตอร์ θ ด้วย $\hat{\theta}$ เมื่อขนาดของกลุ่มตัวอย่างใหญ่ขึ้น ตัวประมาณ $\hat{\theta}$ ที่มีคุณสมบัตินี้ จะประมาณค่าเข้าไปใกล้ค่าพารามิเตอร์มากขึ้นด้วย สามารถเขียนอยู่ในรูปประโยคสัญลักษณ์ทั่วไปได้คือ ถ้า $P(\hat{\theta} \to \theta) \to 1$ เมื่อ $n \to \infty$ แล้ว $\hat{\theta}$ จะเรียกว่าเป็นตัวประมาณค่าที่มีความสอดคล้องของ θ ซึ่งก็หมายความว่า ความน่าจะเป็นที่ค่า $\hat{\theta}$ จะประมาณเข้าใกล้ θ มีค่าสูงขึ้น เมื่อขนาดกลุ่มตัวอย่าง (n) มีค่าเพิ่มมากขึ้น ๆ คือมีค่าความน่าจะเป็นเข้าใกล้ 1

3. ความมีประสิทธิภาพ (Efficiency) หมายถึง ตัวประมาณค่าหนึ่ง ๆ สามารถประมาณค่าพารามิเตอร์ได้ถูกต้องแม่นยำ (Accuracy) เพียงใด ซึ่งเกณฑ์ที่ใช้พิจารณาความมีประสิทธิภาพของตัวประมาณค่าก็คือ ค่าความแปรปรวนของตัวประมาณที่เปรียบเทียบกับกลุ่มของตัวประมาณที่ไม่เอนเอียงด้วยกัน

โดยทั่วไปจะนิยามประสิทธิภาพของตัวประมาณค่าใด ๆ ว่าเป็นอัตราส่วนระหว่างค่าความแปรปรวนของตัวประมาณที่มีประสิทธิภาพนั้นกับค่าความแปรปรวนของตัวประมาณตัวอื่น ๆ กล่าวคือถ้าค่าความแปรปรวนของ $\hat{\theta}_i$ ($Var(\hat{\theta}_i)$) น้อยกว่าค่าความแปรปรวนของ $\hat{\theta}_j$ ($Var(\hat{\theta}_j)$) เมื่อทั้ง $\hat{\theta}_i$ และ $\hat{\theta}_j$ เป็นตัวประมาณค่าที่ไม่เอนเอียงแล้ว จะได้ว่า $\hat{\theta}_i$ เป็นตัวประมาณค่าที่มีประสิทธิภาพของ $\hat{\theta}_i$ หรือ $\hat{\theta}_j$ จะดีกว่า $\hat{\theta}_j$ ใด ๆ

$$\text{เมื่อ} \quad E_f = \frac{Var(\hat{\theta}_i)}{Var(\hat{\theta}_j)} \quad \text{และ} \quad 0 \leq E_f \leq 1$$

และประสิทธิภาพของตัวประมาณค่า $\hat{\theta}_i$ จะเรียกว่า ตัวประมาณที่มีความแปรปรวนน้อยที่สุดของ θ (minimum variance estimator)

4. ความพอเพียง (Sufficiency) หมายถึง ตัวประมาณค่า $\hat{\theta}$ จะเป็นตัวประมาณค่าที่มีความพอเพียง ถ้ามันให้สารสนเทศที่ก่อเกิดประโยชน์ได้ทั้งหมดที่ต้องการเกี่ยวกับพารามิเตอร์ที่ต้องการประมาณ เช่น $\bar{X}$ เป็นตัวประมาณที่มีความพอเพียงของ μ ก็หมายความว่าไม่มีตัวประมาณค่าของ μ ตัวอื่น ตัวอย่างเช่น มัธยฐาน (Median) ที่จะสามารถให้ข่าวสารเกี่ยวกับ μ เพิ่มขึ้นได้อีก

สำหรับคุณสมบัติของวิธีประมาณค่าที่ดีทั้ง 4 ประการดังกล่าว เป็นเกณฑ์ที่ใช้ในการตัดสินใจเลือกวิธีประมาณค่าทางทฤษฎีกรณีที่มีวิธีประมาณค่าอยู่หลายวิธี และวิธีนั้นจะมีคุณสมบัติข้อใดบ้างสามารถใช้หลักการพิสูจน์ให้เห็นจริงได้

<u>การแจกแจงแบบปกติสองตัวแปร</u> (Bivariate Normal Distribution)

ถ้า X และ Y เป็นตัวแปรสุ่มแบบต่อเนื่อง มีการแจกแจงแบบปกติสองตัวแปร ฟังก์ชั่นความน่าจะเป็นของตัวแปรทั้งสอง เขียนเป็นสมการได้ดังนี้

$$f(X, Y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\{ -\frac{1}{2(1-\rho^2)}[\frac{(X-\mu_1)^2}{\sigma_1} - 2\rho \frac{(X-\mu_1)}{\sigma_1}\frac{(Y-\mu_2)}{\sigma_2} + \frac{(Y-\mu_2)^2}{\sigma_2}]\}$$

เมื่อ $-\infty < X < \infty$, $-\infty < Y < \infty$

การแจกแจงแบบปกติสองตัวแปรตามสมการดังกล่าว จะประกอบด้วยพารามิเตอร์ μ_1 , μ_2 , σ_1 , σ_2 และ ρ โดยที่ $-\infty < \mu_1 < \infty$, $-\infty < \mu_2 < \infty$, $\sigma_1 > 0$, $\sigma_2 > 0$ และ $-1 < \rho < 1$ จะได้ X และ Y เป็นตัวแปรที่มีการแจกแจงแบบปกติสองตัวแปรที่มี

ค่าเฉลี่ย $\underset{\sim}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

และ เมตริกความแปรปรวนร่วม $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

โดยที่ ρ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X และ Y

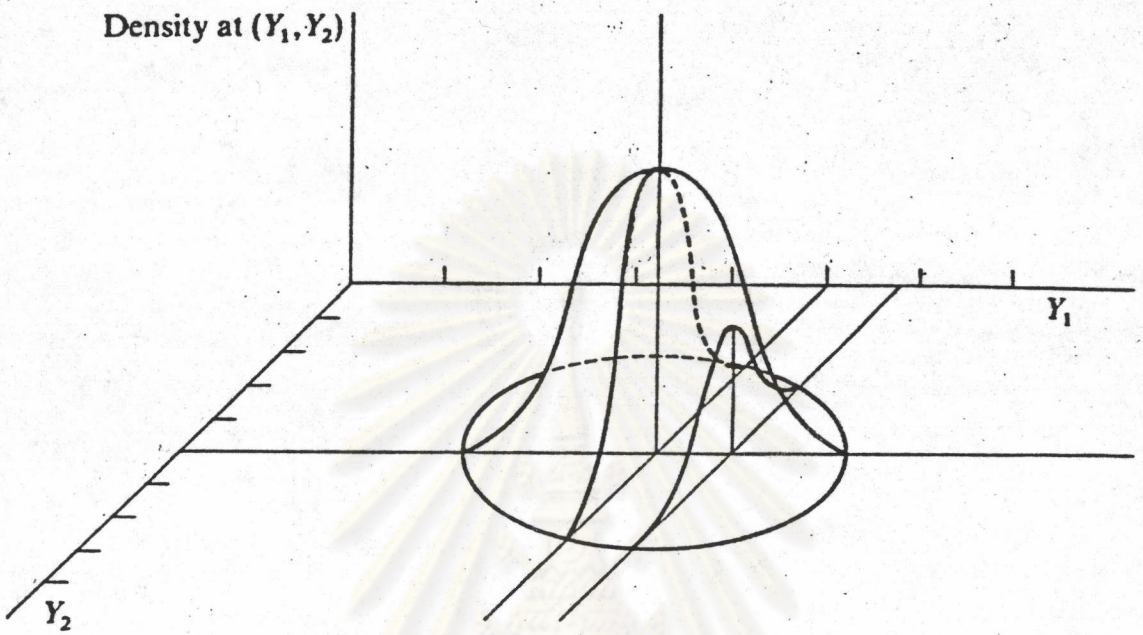
μ_1 คือ ค่าเฉลี่ยของตัวแปร X

μ_2 คือ ค่าเฉลี่ยของตัวแปร Y

σ_1^2 คือ ค่าความแปรปรวนของตัวแปร X

σ_2^2 คือ ค่าความแปรปรวนของตัวแปร Y

ฟังก์ชั่นความน่าจะเป็นของตัวแปร X และ Y ที่มีค่าเฉลี่ยและเมตริกความแปรปรวนร่วมดังกล่าวสามารถเขียนแสดงได้ ดังแผนภาพที่ 1



แผนภาพที่ 1 การแจกแจงแบบปกติสองตัวแปร

ถ้าตัวแปรสุ่ม X และ Y มีการแจกแจงแบบปกติสองตัวแปร การแจกแจงแบบมีเงื่อนไขของตัวแปร X เมื่อกำหนดค่า $Y = y$ หรือเขียนแทนด้วยสัญลักษณ์ $X|Y = y$ จะมีลักษณะเป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $\mu_1 + (\rho\sigma_1|\sigma_2)(Y - \mu_2)$ และค่าความแปรปรวนเท่ากับ $\sigma_1^2(1 - \rho^2)$ และการแจกแจงแบบมีเงื่อนไขของตัวแปร Y เมื่อกำหนดค่า $X = x$ ซึ่งเขียนแทนด้วยสัญลักษณ์ $Y|X = x$ จะมีลักษณะการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $\mu_2 + (\rho\sigma_2|\sigma_1)(X - \mu_1)$ และค่าความแปรปรวนเท่ากับ $\sigma_2^2(1 - \rho^2)$ (Brownlee 1965: 404) และเพราะว่า (X , Y) มีการแจกแจงแบบปกติสองตัวแปร ที่มีค่าเฉลี่ย μ_1 , μ_2 ค่าความแปรปรวน σ_1^2 , σ_2^2 และค่าสัมประสิทธิ์สหสัมพันธ์ ρ จะสรุปได้ว่า $Z_1 = \frac{X - \mu_1}{\sigma_1}$ และ $Z_2 = \frac{Y - \mu_2}{\sigma_2}$ จะมีการแจกแจงแบบปกติสองตัวแปรที่มีค่าเฉลี่ย $\mu_{Z_1} = \mu_{Z_2} = 0$, $\sigma_{Z_1}^2 = \sigma_{Z_2}^2 = 1$ และ $\rho(Z_1 , Z_2) = \rho$ (กรรณิกา เลียงเจริญสิทธิ์ 2527: 10)

จากสมการ $\mu_{Y|X=x} = \mu_2 + (\rho\sigma_2/\sigma_1)(X - \mu_1)$

และ $\sigma^2_{Y|X=x} = \sigma^2_2(1 - \rho^2)$

จะได้ $Y \sim N(\sigma^2_2(1 - \rho^2) , \mu_2 + (\rho\sigma_2/\sigma_1)(X - \mu_1))$

จากแผนฟังการสร้างตัวแปรสุ่มปกติสองตัวแปร ค่าตัวแปร Y จะขึ้นอยู่กับค่าตัวแปร X สามารถสร้างค่าตัวแปรทั้งสองได้จาก สมการต่อไปนี้

$$Y = \mu_2 + (\rho\sigma_2/\sigma_1)(X - \mu_1) + \sqrt{\sigma^2_2 (1 - \rho^2)} \cdot Y1$$

เมื่อ Y1 คือ ค่าตัวแปรอิสระจากค่าตัวแปร X มี $E(Y1) = 0$

และ $Var(Y1) = 1$

งานวิจัยที่เกี่ยวข้อง

ชะไมพร ธรรมวัฒน์ไพศาล (2522: 124-126) ได้ศึกษาเกี่ยวกับวิธีการประมาณค่าข้อมูลที่สูญหายในการวิเคราะห์การถดถอยวิธีต่าง ๆ 6 วิธีด้วยกัน คือ วิธีกำลังสองน้อยที่สุด วิธีอันดับศูนย์หรือวิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง วิธีอันดับศูนย์ดัดแปลง วิธีถดถอยอันดับหนึ่งหรือวิธีใช้สมการถดถอย วิธีถดถอยสองชั้น และวิธีผสมหรือวิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยจากกลุ่มตัวอย่างและสมการถดถอย โดยใช้ข้อมูลที่รวบรวมมาได้และกระทำให้สูญหายไปโดยวิธีสุ่ม ใช้เกณฑ์เปรียบเทียบคือค่าความแปรปรวนร่วมที่สามารถอธิบายได้หรือ (R^2) (a coefficient of determination) ซึ่งค่านี้สามารถใช้เป็นดัชนีในการตัดสินใจได้ว่าวิธีประมาณค่าใดสามารถประมาณค่าได้ใกล้เคียงกับค่าที่สูญหายมากกว่ากัน ถ้า (R^2) มีค่าเพิ่มมากขึ้น (Ozer 1985: 307-308) จากงานวิจัยนี้สรุปได้ว่าวิธีที่ให้ค่า (R^2) สูงกว่าวิธีอื่นมีอยู่ 3 วิธีด้วยกัน เรียงตามลำดับจากมากไปน้อย คือ วิธีถดถอยสองชั้น วิธีถดถอยอันดับหนึ่งหรือวิธีใช้สมการถดถอย และวิธีกำลังสองน้อยที่สุด

พรศิริ หมื่นไชยศรี (2529: 75-76) ได้ศึกษาเกี่ยวกับการเปรียบเทียบวิธีประมาณค่าข้อมูลที่สูญหายในการวิเคราะห์ตัวแปรพหุคูณ 4 วิธี คือ วิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง วิธีวิเคราะห์ความถดถอยพหุคูณเชิงเส้น วิธีวิเคราะห์ความถดถอยพหุคูณเชิงเส้นดัดแปลง และวิธีวิเคราะห์ส่วนประกอบหลัก ทำการศึกษาโดยใช้เทนิคมอนติคาร์โลซิมูเลชั่น และใช้เกณฑ์ในการเปรียบเทียบ คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (mean square error) และคะแนนรวมจากการถ่วงน้ำหนักของทั้ง 4 วิธีที่ได้ลำดับ 1, 2, 3 และ 4

เมื่อเรียงลำดับค่าเฉลี่ยความคลาดเคลื่อนกำลังสองจากน้อยไปหามาก โดยศึกษาที่ขนาดกลุ่มตัวอย่าง 30, 50, 70, 100 และ 200 จำนวนตัวแปรเท่ากับ 3, 5, 7 และ 10 ตัวแปร ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร (ρ) เท่ากับ 0.1 , 0.2 ..., 0.9 และกำหนดสัดส่วนข้อมูลที่สูญหายของแต่ละตัวแปรมีค่าใกล้เคียงกันคือ 10 % จากการศึกษาพบว่าวิธีประมาณค่าข้อมูลที่สูญหายในการวิเคราะห์ตัวแปรพหุคูณทั้ง 4 วิธี ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองไม่แตกต่างกันอย่างมีนัยสำคัญที่ระดับ .05 ไม่ว่าจะเป็นสถานการณ์ใดก็ตามที่มีข้อมูลสูญหายเกิดขึ้น

ชานและดันน (Chan and Dunn 1972: 473-477) ใช้เทคนิคมอนติคาร์โลซิมูเลชั่นทำการศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลที่สูญหายในการวิเคราะห์จำแนกประเภท (Discriminant Analysis) จำนวน 5 วิธีด้วยกัน คือ 1. ศึกษาเมื่อไม่มีข้อมูลสูญหายเลย 2. วิธีตัดตัวอย่างที่มีข้อมูลสูญหายออก 3. วิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง 4. วิธีใช้สมการถดถอย และ 5. วิธีวิเคราะห์ส่วนประกอบหลัก (principal component) โดยศึกษาข้อมูลที่มีลักษณะการแจกแจงแบบปกติสองตัวแปร และประชากรทั้งสองกลุ่มมีความแปรปรวนเท่ากัน ใช้เกณฑ์ในการพิจารณาเปรียบเทียบคือ ร้อยละของการจำแนกข้อมูลผิด พบว่าโดยทั่ว ๆ ไป ค่า R^2 (Coefficient of determination) ของวิธีใช้สมการถดถอย จะสูงกว่าวิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง วิธีวิเคราะห์ส่วนประกอบหลัก และวิธีตัดตัวอย่างที่มีข้อมูลสูญหายออก และถ้าจำนวนตัวแปรเพิ่มขึ้นวิธีใช้สมการถดถอยจะให้ค่า R^2 เพิ่มสูงขึ้น

อะฟีฟี และอีลาสฮอฟฟ์ (Afifi and Elashoff 1967: 18-28) ได้ทำการศึกษาเปรียบเทียบวิธีประมาณค่า คือ วิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง (The Zero Order method) วิธีใช้ค่าเฉลี่ยดัดแปลง (A modified Zero Order method) และวิธีผสม หรือวิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยจากกลุ่มตัวอย่างและสมการถดถอย (Mixed method) โดยเปรียบเทียบ

ประสิทธิภาพของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของแต่ละวิธีเทียบกับวิธีกำลังสองน้อยที่สุดในกลุ่มตัวอย่างขนาดใหญ่คือมากกว่าหรือเท่ากับ 200 พบว่าเมื่อค่าสัมประสิทธิ์สหสัมพันธ์และจำนวนการสูญหายเท่ากับประสิทธิภาพของวิธี Zero order และวิธี Modified Zero order จะลดลงเมื่อกลุ่มตัวอย่างเพิ่มขึ้น ถ้ากำหนดขนาดกลุ่มตัวอย่างให้คงที่โดยเปลี่ยนแปลงจำนวนการสูญหายของตัวแปรอิสระ (m_x) และจำนวนการสูญหายของตัวแปรตาม (m_y) พบว่าวิธี Zero order มีประสิทธิภาพสูงที่สุดเมื่อการสูญหายของตัวแปรอิสระน้อยกว่าตัวแปรตาม ($m_x < m_y$) ที่ค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์เล็กน้อย ($|\rho|$) ซึ่งตรงข้ามกับ Mixed method จะมีประสิทธิภาพสูงที่สุดเมื่อการสูญหายของตัวแปรตามน้อยกว่าการสูญหายของตัวแปรอิสระ ($m_x > m_y$) ที่ค่าสัมบูรณ์ของสัมประสิทธิสหสัมพันธ์อยู่ในช่วงกลาง ๆ

ฟิงค์เบนเนอร์ (Finkbeiner 1979: 416-420) ใช้เทคนิคมอนติคาร์โลซิมูเลชันทำการศึกษาเปรียบเทียบความแม่นยำของวิธีประมาณค่าข้อมูลที่สูญหายในการวิเคราะห์ตัวแปรพหุคูณ 6 วิธีด้วยกันคือ 1. วิธี maximum likelyhood (ML) 2. วิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง (MR) 3. วิธีใช้เฉพาะตัวอย่างที่ไม่มีข้อมูลสูญหาย (CD) 4. วิธีใช้สมการถดถอย (REG) 5. วิธีวิเคราะห์องค์ประกอบหลัก (PC) และ 6. วิธีใช้เฉพาะคู่ตัวอย่างที่สมบูรณ์ (CPO) ใช้เกณฑ์ในการเปรียบเทียบต่อไปนี้คือ ใช้ค่าเฉลี่ยจากกลุ่มตัวอย่างและการกระจายของค่าพารามิเตอร์ และใช้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธีที่มีค่าน้อยที่สุด โดยศึกษาที่ขนาดกลุ่มตัวอย่างเท่ากับ 64 จำนวนการสูญหายของข้อมูล 2 รูปแบบ (pattern) ทำการทดลองซ้ำ 50 ครั้ง จากการศึกษาพบว่า ค่าเฉลี่ยจากกลุ่มตัวอย่างและค่าส่วนเบี่ยงเบนมาตรฐานของทุกวิธีไม่แตกต่างกัน แต่ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองแตกต่างกันเรียงลำดับจากน้อยไปมาก คือ ML , MR , CPO และ REG ซึ่งสรุปได้ว่าวิธีที่สามารถประมาณค่าข้อมูลที่สูญหายไปในรูปแบบการวิเคราะห์ตัวแปรพหุคูณ และมีความแม่นยำอยู่ในเกณฑ์ดีมากมีอยู่สามวิธีคือวิธี maximum likelyhook มีความแม่นยำในการประมาณมากที่สุด รองลงมาคือ วิธีแทนด้วยค่าเฉลี่ย (MR) และวิธีใช้เฉพาะคู่ตัวอย่างที่สมบูรณ์ (CPO) ส่วนวิธีใช้สมการถดถอย (REG) ยังอยู่ในขั้นต่ำกว่าเกณฑ์ที่กำหนด