บทที่ 2

สถิติที่ใช้ในการวิจัย

ในการวิจัยครั้งนี้สนใจการประมาณค่าเฉลี่ยของประชากร(Population mean) ของตัวแปรที่สนใจ โดยการสุ่มอย่างง่ายโดยใช้วิธีการดังต่อไปนี้

วิธีการประมาณค่าพารามิเตอร์เฉพาะข้อมูลที่เก็บรวบรวมมาได้หรือค่าเฉลี่ยจากการตอบกลับ

การประมาณโดยวิธีนี้จะประมาณเฉพาะข้อมูลที่เก็บมาได้ โดยไม่คำนึงถึงข้อมูลที่ขาดหายไป ดังนั้นตัวประมาณค่าเฉลี่ยประชากรมีรูปแบบดังนี้

การประมาณค่าเฉลี่ยของประชากร(Population mean)

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

ประมาณโดยใช้ค่าเฉลี่ยตัวอย่าง

$$\overline{y} = \frac{\sum_{i=1}^{n_r} y_i}{n_r}$$

เมื่อกำหนด

y_i = ค่าของลักษณะที่ต้องการศึกษาของหน่วยที่ i

N = จำนวนหน่วยทั้งหมดในประชากร

n_r = จำนวนหน่วยตัวอย่างตอบสัมภาษณ์

การประมาณค่าพารามิเตอร์โดยใช้เทคนิคของ **Hansen&Hurwitz**

การประมาณโดยวิธีนี้โดยการเลือกตัวอย่างย่อย(subsample)มาจำนวนหนึ่งจากตัวอย่างทั้งหมดที่ไม่ตอบสัมภาษณ์แล้วสัมภาษณ์ตัวอย่างย่อยที่เลือกมาให้ได้ครบทุกตัวอย่าง โดยกำหนดสัญลักษณ์ดังนี้

n = จำนวนหน่วยตัวอย่างที่เลือกมาโดยสุ่มจากประชากร

n_1 = จำนวนหน่วยตัวอย่างที่ตอบสัมภาษณ์

n_2 = จำนวนหน่วยตัวอย่างที่ไม่ตอบสัมภาษณ์

n_2' = จำนวนหน่วยตัวอย่างย่อยที่เลือกขึ้นมาโดยสุ่มตัวอย่างจากที่ไม่ตอบสัมภาษณ์

จะประมาณค่าเฉลี่ยของประชากร โดยใช้ค่าเฉลี่ยถ่วงน้ำหนักของค่าเฉลี่ยที่ตอบสัมภาษณ์ และค่าเฉลี่ยจากตัวอย่างย่อยที่เลือกมาจากตัวอย่างที่ไม่ตอบสัมภาษณ์ โดยมีจำนวนตัวอย่างที่ตอบสัมภาษณ์และไม่ตอบสัมภาษณ์เป็นตัวถ่วงน้ำหนักกล่าวคือ

$$\overline{y} = \frac{n_1 \overline{y}_1}{n} + \frac{n_2 \overline{y}_2}{n}$$

โดยที่

$$\overline{y}_1 = \frac{\sum_{i=1}^{n_1} y_{1i}}{n_1}$$

$$\overline{y}_2 = \frac{\sum_{i=1}^{n_2'} y_{2i}}{n_2'}$$

ซึ่งตัวประมาณที่ได้นี้เป็นตัวประมาณที่ไม่เอนเอียง

การประมาณค่าเฉลี่ยประชากรเมื่อเกิดปัญหาการไม่ตอบสัมภาษณ์เสนอโดย Phillip S.Kott

ในการประมาณค่าเฉลี่ยของประชากรด้วยตัวอย่างขนาด n ถ้ากำหนด

p_i = ความน่าจะเป็นของการถูกเลือก สามารถเขียนตัวประมาณได้ในรูปแบบของ

$$\overline{Y} = \frac{\sum_{i=1}^{N} w_i y_i}{N}$$

ซึ่ง $w_i = \frac{1}{p_i}$ และ Phillip S.Kott ได้เสนอรูปแบบตัวประมาณดังนี้

$$\overline{y} = \frac{(\sum_{i \in R} w_i y_i + \sum_{i \in M} w_i x_i b^*)}{N}$$

M แทนกลุ่มของหน่วยตัวอย่างซึ่งไม่ตอบกลับ

$b^* = (X_R{}' D_R^{-1} Q_R [I_R - C_R] X_R)^{-1} X_R{}' D_R^{-1} Q_R [I_R - C_R] y_R$

I_R เป็นเมตริกซ์เอกลักษณ์ขนาด $r \times r$

C_R เป็นค่าใน diagonal matrix ที่เหมาะกับค่า c_i

c_i คือค่าความน่าจะเป็นของการตอบกลับ

β เป็นเวกเตอร์คอลัมภ์ขนาด k

ε_i เป็นตัวแปรสุ่มที่มีค่าเฉลี่ยเท่ากับ 0

ตัวประมาณ $\overline{y}$ ที่ได้เป็นตัวประมาณไม่เอนเอียงเมื่อคำนึงถึงการสุ่มตัวอย่างตั้งแต่เริ่มแรก นั่นคือ

$$E_P[E_M(\overline{y} - \mu)] = 0$$

ค่าประมาณของความแปรปรวน ในการประมาณค่าความแปรปรวนของตัวประมาณ $\overline{y}$ จะเขียนตัวประมาณในรูปแบบถ่วงน้ำหนักคือ

$$\overline{y} = \frac{\sum_{R} a_i^* y_i}{N}$$

เมื่อ $a_i^* = q_i\{c_i + (\sum_{M} w_j x_j) \times (X_R' D_R^{-1} Q_R [I_R - C_R] X_R)^{-1} x_i' d_i^{-1}(1-c_i)\}$

ในส่วนของการตอบกลับ b^* มีค่าจำกัดคือ B และกำหนดให้ $e_i = y_i - x_i\beta$

จากตัวประมาณซึ่งประกอบด้วย 2 ส่วนคือส่วนของการตอบกลับและการไม่ตอบกลับ ดังนั้นความแปรปรวนจึงประกอบด้วยเทอมแรกเป็นของ $\frac{\sum_{R} a_i^* e_i^*}{N}$ และความแปรปรวนเทอมที่สองเป็นของ $\frac{\sum_{s} w_i x_i B^*}{N}$ และความแปรปรวนร่วมของ $\frac{\sum_{R} a_i^* e_i^*}{N}$ และ $\frac{\sum_{s} w_i x_i B^*}{N}$

ดังนั้นตัวประมาณสำหรับเทอมแรกคือ

$$v_1 = \{\sum_{i\in R}\sum_{j\in R}([\pi_{ij} - \pi_i \pi_j / \pi_{ij})(a_i^* e_i^*) a_j^* e_j^*)\} / N$$

เมื่อ $e_i^* = y_i - x_i b^*$

π_i เป็นความน่าจะเป็นที่หน่วยที่ i ถูกเลือกอยู่ในตัวอย่าง

π_{ij} เป็นความน่าจะเป็นที่หน่วยที่ i และ j ถูกเลือกอยู่ในตัวอย่าง

ความแปรปรวนสำหรับเทอมที่สองคือ

$$v_2 = \frac{\sum_{i\in S}\sum_{j\in S}([p_{ij} - p_i p_j] / p_{ij})(w_i x_i b^*)(w_j x_j b^*)}{N}$$

เมื่อ p_i เป็นความน่าจะเป็นที่หน่วยที่ i ถูกเลือกอยู่ในตัวอย่าง

p_{ij} เป็นความน่าจะเป็นที่หน่วยที่ i และ j ถูกเลือกอยู่ในตัวอย่าง

และเทอมความแปรปรวนร่วมคือ

$$v_3 = \frac{2\sum_{i\in R}\sum_{j\in R}([p_{ij} - p_i p_j]/\pi_{ij}])(w_i x_i b^*)(w_j e_j^*)}{N}$$

เพราะว่า $\frac{\sum_R a_i^* e_i^*}{N}$ ภายใต้รูปแบบการตอบกลับประมาณด้วย $\frac{\sum_s w_i e_i^*}{N}$ สังเกตว่า $E_M(v_3) = 0$ ดังนั้นความแปรปรวนของตัวประมาณคือ $v(\bar{y}) = v_1 + v_2 + v_3$

วิธีการประมาณค่าพารามิเตอร์ที่ดัดแปลงโดยผู้วิจัย

วิธีการประมาณนี้ได้จากการดัดแปลงการหาค่า nonresponse จากวิธีของ Isaac Olayiwola Oshungade ได้เสนอวิธีการโดยอาศัยหลักการของเบยส์ โดยวิธีการนำ ตัวแปรช่วยเข้ามาเกี่ยวข้อง โดยกำหนดเป็นตัวแปร x เมื่อตัวแปรที่สนใจคือตัวแปร y นอกจากนั้นยังกำหนดข้อมูลทั้ง x และ y เป็นช่วงหรือกลุ่มด้วย

ข้อตกลงเบื้องต้น

i)ตัวแปร $X = (x_1, x_2, x_3, \ldots x_j)$เป็นอิสระกันและความน่าจะเป็นของการเกิดขึ้นของตัวแปร x_j แต่ละช่วงของตัวแปร y ที่สนใจคือ

$$p(X / y_i) = p(x_1 / y_i) p(x_2 / y_i) \ldots p(x_j / y_i)$$

ii)prior probability ของการประมาณช่วง $p(y_i)$ และ $p(x_j / y_i)$ ทราบค่าตลอดจนสัดส่วนของแต่ละชั้นของ y_i และค่าของ x_j ในแต่ละกลุ่มของ y_i ด้วย

iii) y_i จะมีค่าอยู่เพียงกลุ่มเดียว และ $p(y_i) = 1$ ความน่าจะเป็นต่างๆคำนวนได้ดังนี้คือ

$p(y_i)$ คือสัดส่วนของจำนวนซึ่งตกอยู่ในชั้น i ของy

$p(x_j / y_i)$ คือสัดส่วนของจำนวนของx_j ซึ่งตกอยู่ในชั้นของ y_i และ $p(y_i / x)$คำนวนได้จาก

$$p(y_i / x) = \frac{p(y_i)\prod_{j=1}^{m} p(x_j / y_i)}{\sum [p(y_i)\prod_{j=1}^{m} p(x_j / y_i)}$$

(i คือจำนวนชั้นของ y) เมื่อคำนวน $p(y_i / x)$ครบทุกชั้น เลือกชั้นให้ความน่าจะเป็นสูงสุด เพราะฉะนั้น ค่าy_i ที่ไม่ตอบ สัมภาษณ์ควรจะอยู่ในช่วงที่ i หรือมีค่าอยู่ในกลุ่ม i นั่นเอง

จากวิธีการของ Isaac Olayiwola oshungade สามารถทำนายค่า NOnresponse ออกมาในลักษณะเป็นชั้นหรือช่วงของข้อมูลทำให้สามารถทราบได้ว่าค่าของตัวแปรที่สนใจซึ่งเกิดการไม่ตอบกลับอยู่ในช่วงใด ดังนั้นผู้วิจัยได้เสนอการนำเสนอค่าทำนายออกมาในลักษณะเป็นจุดเพื่อที่จะสามารถนำค่านี้ไปแทนหน่วยที่ไม่ตอบสัมภาษณ์ในตัวแปรที่สนใจ โดยดำเนินการต่อจากวิธีการของ Isaac Olayiwola Oshungade ดังนี้

1 การประมาณค่าที่ไม่ตอบกลับด้วยสมการถดถอย

i)ในแต่ละชั้นหรือช่วงของy คือแต่ละy_i สร้างสมการถดถอยระหว่างตัวแปร y กับ x_jดังนั้นจะได้จำนวนสมการถดถอยเท่ากับจำนวนชั้นของ y

ii) เมื่อได้ความน่าจะเป็นของ y_i ตกอยู่ในชั้นใด จะใช้สมการถดถอยในชั้นนั้นพยากรณ์ค่าของ y_{ik} (k แทนค่าสังเกตที่ k)ที่ไม่ตอบกลับนั้น

จากวิธีการของIsaac Olayiwola Oshungade ผู้วิจัยจึงเสนอตัวประมาณ μ คือ

$$\bar{y} = \frac{\sum_{i \in R} w_i y_i + \sum_{i \in M} w_i x_i b_i}{N}$$

เมื่อ w_i เป็นความน่าจะเป็นที่จะตกเป็นตัวอย่าง

R แทนกลุ่มของการตอบกลับ

M แทนกลุ่มของการไม่ตอบกลับ

b_i เป็นค่าสัมประสิทธิ์การถดถอยในชั้นที่ i โดยที่ $y_i = Xb_i$ เป็นค่าทำนายสำหรับหน่วยที่ไม่ตอบกลับ

2 การประมาณค่าที่ไม่ตอบกลับด้วยค่าเฉลี่ยของกลุ่ม

i) วิธีนี้เริ่มจากเมื่อแบ่งกลุ่มหรือชั้นในตัวแปรที่สนใจได้แล้ว หาค่าเฉลี่ยในแต่ละกลุ่มจากค่าที่ตอบกลับ เมื่อเกิดปัญหาการไม่ตอบกลับหลังจากคำนวณหาค่าความน่าจะเป็นได้ว่าค่าที่ไม่ตอบกลับควรจะตกอยู่ในกลุ่มใด จะใช้ค่าเฉลี่ยในกลุ่มนั้นเข้าไปแทน ดังนั้นตัวประมาณค่าเฉลี่ยจะอยู่ในรูปต่อไปนี้

$$\bar{y} = \frac{\sum_{i \in R} w_i y_i + \sum_{i \in M} w_i \bar{y}_i}{N}$$

เมื่อ w_i เป็นความน่าจะเป็นที่จะตกเป็นตัวอย่าง

R แทนกลุ่มของการตอบกลับ

M แทนกลุ่มของการไม่ตอบกลับ

$\bar{y}_i$ แทนค่าเฉลี่ยของกลุ่ม

การประมาณค่ารวมและสัดส่วน

1. การประมาณค่ารวม

ในการสุ่มตัวอย่างอย่างง่าย เราจำเป็นต้องทราบจำนวนหน่วยตัวอย่าง เราจึงประมาณค่ารวมได้โดยอาศัยตัวประมาณค่าเฉลี่ยเป็นหลัก กล่าวคือ

$$\hat{Y} = N\bar{y}$$

เมื่อ $\hat{Y}$ เป็นค่ารวมของสิ่งที่สนใจจากประชากร

N ขนาดของประชากร

$\bar{y}$ ค่าเฉลี่ยจากตัวอย่าง

2. การประมาณสัดส่วน(Proportion)

เนื่องจากลักษณะหรือสิ่งที่เราสนใจจะทราบเกี่ยวกับแต่ละหน่วยในประชากรอาจไม่ใช่ค่าที่วัดได้เป็นปริมาณเสมอไป แต่อาจเป็นลักษณะหรือคุณภาพแต่ละหน่วย เช่นหน่วยที่ i มีลักษณะที่เราสนใจหรือไม่จะเห็นว่าลักษณะคำตอบจะเป็นไปได้ 2 ลักษณะ ดังนั้นเราสามารถแบ่งประชากรออกเป็น 2 ส่วนตามลักษณะที่ได้

กำหนด A เป็นจำนวนหน่วยในประชากรที่มีลักษณะที่ต้องการ

a เป็นจำนวนหน่วยตัวอย่างที่มีลักษณะที่ต้องการ

แล้ว สัดส่วนของประชากรที่มีลักษณะนี้คือ $P = \frac{A}{N}$

และถ้าเราเลือกตัวอย่างสุ่มขนาด n เราอาจประมาณสัดส่วนนี้ได้ด้วยค่า $P = \frac{a}{n}$

โดยใช้คุณสมบัติของการประมาณค่าเฉลี่ยเมื่อพิจารณาค่าของ y_i ดังนี้

กำหนดให้ $y_i = 1$ ถ้ามีลักษณะที่ต้องการ

$y_i = 0$ ถ้าไม่มีลักษณะที่ต้องการ

ดังนั้นตัวประมาณสัดส่วนที่ได้จะมีลักษณะเดียวกับตัวประมาณค่าเฉลี่ยประชากรคือ

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$