บทที่ 1

บทนำ

ที่มาและความสำคัญของปัญหา

ในการศึกษาการประมาณค่าพารามิเตอร์โดยทั่วไป สามารถทำการประมาณได้ใน 2 รูปแบบ คือการประมาณค่าแบบจุด (Point estimation) และการประมาณแบบช่วง (Interval estimation) สำหรับการประมาณแบบจุดเป็นการประมาณค่าพารามิเตอร์ด้วยค่าๆหนึ่ง ซึ่งการประมาณค่าแบบจุด ค่าประมาณที่ได้จะคลาดเคลื่อนไปจากค่าพารามิเตอร์เพียงใด ขึ้นอยู่กับการเลือกใช้ตัวประมาณที่เหมาะสม ส่วนการประมาณค่าแบบช่วงเป็นการประมาณที่จะให้ช่วงๆหนึ่งซึ่งมีคุณสมบัติว่าค่าที่แท้จริงของประชากรจะอยู่ในช่วงที่ประมาณได้ด้วยความเชื่อมั่นระดับหนึ่ง โดยที่การประมาณค่าแบบช่วงเป็นการประมาณโดยอาศัยตัวประมาณแบบจุด และการแจกแจงความน่าจะเป็นของตัวประมาณนั้น ซึ่งผลจากการประมาณจะทำให้ผู้วิจัยเชื่อมั่นได้ในระดับหนึ่งว่าช่วงที่ประมาณได้คลุมค่าพารามิเตอร์ที่สนใจศึกษา จะเห็นได้ว่าการประมาณค่าแบบช่วงสามารถบอกขอบเขตของค่าประมาณได้ดีกว่าการประมาณค่าแบบจุด ที่ให้ค่าประมาณที่ได้เป็นเพียงค่าๆหนึ่ง

สำหรับในการวิจัยครั้งนี้ ผู้วิจัยสนใจศึกษาการประมาณค่าแบบช่วงสำหรับผลต่างระหว่างค่าสัดส่วนของสองประชากร ถ้าให้ X_1 และ X_2 เป็นตัวแปรเชิงสุ่มที่เป็นอิสระกัน และต่างมีการแจกแจงแบบทวินามด้วยพารามิเตอร์ n_1, p_1 และ n_2, p_2 ตามลำดับ และ $0 \le p_1, p_2 \le 1$ นั่นคือ X_1 และ X_2 คือจำนวนผลสำเร็จทั้งหมดในการทดลองแบร์นูลลี่ที่เป็นอิสระกัน n_1 และ n_2 ครั้งตามลำดับ การประมาณค่าแบบช่วงสำหรับผลต่างระหว่างค่าสัดส่วนของสองประชากรสามารถสร้างได้ดังนี้

ขณะที่ n_1 และ n_2 มีขนาดใหญ่ จะได้ว่า X_1 และ X_2 จะมีการแจกแจงแบบปกติ N$(n_1p_1, n_1p_1(1-p_1))$ และ N$(n_2p_2, n_2p_2(1-p_2))$ ตามลำดับ และเนื่องจากตัวประมาณแบบจุดของ p_1 และ p_2 ซึ่งเป็นค่าสัดส่วนของประชากรที่ 1 และ 2 (Population proportion) คือค่าสัดส่วนตัวอย่างที่ 1 และ 2 ($\hat{p}_1, \hat{p}_2$: Sample proportion) ตามลำดับ โดยที่ $\hat{p}_1 = \dfrac{X_1}{n_1}$ และ $\hat{p}_2 = \dfrac{X_2}{n_2}$ ต่างเป็น

ตัวสถิติที่เป็นอิสระซึ่งกันและกัน และมีการแจกแจงปกติที่มีค่าเฉลี่ย p_1 และ p_2 ตามลำดับ และค่าความแปรปรวน $p_1(1-p_1)/n_1$ และ $p_2(1-p_2)/n_2$ ตามลำดับ ($p_i \sim N(p_i, p_i q_i/n_i)$ i =1,2)

ดังนั้นผลต่างของค่าสัดส่วนตัวอย่าง $(\hat{p}_1 - \hat{p}_2)$ เป็นตัวประมาณที่ไม่เอนเอียงของ p_1 - p_2 ที่มีค่าเฉลี่ย p_1 - p_2 และความแปรปรวน $\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}$ และตัวแปรสุ่ม $\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}$ จะมีการแจกแจงแบบปกติมาตรฐาน (Standard normal distribution)

จากวิธีการข้างต้น เมื่อกำหนดระดับนัยสำคัญ α สามารถหาช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับค่าผลต่างระหว่างสัดส่วนของประชากรสองชุด (p_1 - p_2) ได้เป็น

$$(\hat{p}_1-\hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

ซึ่ง $Z_{\alpha/2}$ คือเปอร์เซ็นไทล์ที่ $(1-\alpha/2)100$ ของ N(0,1) รูปแบบของวิธีการประมาณข้างต้นเป็นรูปแบบอย่างง่าย เนื่องจากสามารถประมาณการแจกแจงความน่าจะเป็นแบบทวินามได้ด้วยการแจกแจงแบบปกติ แต่การแจกแจงความน่าจะเป็นแบบทวินาม เป็นการแจกแจงของตัวแปรสุ่มชนิดไม่ต่อเนื่อง (Discrete random variable) ในขณะที่การแจกแจงปกติเป็นการแจกแจงชนิดที่ต่อเนื่อง (Continue random variable) ดังนั้นในการประมาณค่าสัดส่วนประชากรแบบช่วงด้วยการแจกแจงแบบปกติ จึงควรปรับช่วงของความน่าจะเป็นให้ได้ค่าความน่าจะเป็นใกล้ค่าจริง หรือมีความคลาดเคลื่อนน้อย ค่าที่ใช้ปรับแก้ไขเพื่อความต่อเนื่อง (Continuity correction) ในการศึกษาครั้งนี้ ผู้วิจัยสนใจที่จะศึกษาการประมาณค่าแบบช่วงสำหรับผลต่างของสัดส่วนของสองประชากรด้วยวิธีการประมาณด้วยการแจกแจงแบบปกติ และใช้ค่าปรับแก้เพื่อความต่อเนื่องต่างๆ ว่าวิธีการประมาณแบบช่วงวิธีใดจะให้ค่าช่วงความเชื่อมั่นที่มีค่าสัมประสิทธิ์ความเชื่อมั่นไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่ต้องการ และให้ค่าความยาวช่วงต่ำที่สุดในกลุ่มที่ศึกษาเปรียบเทียบ

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบวิธีการประมาณแบบช่วงสำหรับผลต่างระหว่างค่าสัดส่วนของสองประชากร ด้วยวิธีต่อไปนี้

1. วิธีการประมาณอย่างง่าย (Classical Method)
2. วิธีการประมาณโดยใช้ค่าปรับแก้เพื่อความต่อเนื่องของเยจส์ (Yates)

3. วิธีการประมาณโดยใช้ค่าปรับแก้เพื่อความต่อเนื่องของฮอคก์ และแอนเดอร์สัน (Hauck and Anderson)

4. วิธีการประมาณโดยใช้ค่าปรับแก้เพื่อความต่อเนื่องของเพสกัน (Peskun)

การเปรียบเทียบจะเปรียบเทียบจากค่าสัมประสิทธิ์ความเชื่อมั่นของช่วงความเชื่อมั่น และค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นที่คำนวณได้จากแต่ละวิธี ที่ระดับความเชื่อมั่น 3 ระดับ คือ 90%, 95% และ 99%

สมมติฐานของการวิจัย

ในกรณีตัวอย่างขนาดเล็กวิธีการประมาณของเพสกัน (Peskun) จะให้ค่าสัมประสิทธิ์ความเชื่อมั่นไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด และให้ค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นต่ำกว่าการประมาณอื่น ๆ

วิธีการประมาณด้วยการใช้ค่าปรับแก้ไขเพื่อความต่อเนื่อง จะให้ระดับความเชื่อมั่นใกล้เคียงกับวิธีที่ประมาณอย่างง่าย เมื่อขนาดตัวอย่างทั้งสองมีขนาดใหญ่พอ

ข้อตกลงเบื้องต้น

1. การแจกแจงทวินามของแต่ละประชากรเป็นอิสระกัน
2. ขนาดตัวอย่าง n_1, n_2 เป็นพารามิเตอร์ที่ทราบค่า
3. ค่าสัดส่วนความสำเร็จของแต่ละประชากร (p_1,p_2) เป็นพารามิเตอร์ที่ไม่ทราบค่า
4. ค่าสัมบูรณ์ของผลต่างระหว่างค่าสัดส่วนของสองประชากร $|p_1-p_2|$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

ขอบเขตของการวิจัย

1. กำหนดขนาดตัวอย่าง n_1 และ n_2 โดยศึกษาในกรณีที่ $n_1 = n_2$ มีค่าเท่ากับ 10, 20, 25, 30, 35, 40, 50, 60, 70, 80 (และได้ศึกษากรณี $n_1 \neq n_2$ ไว้ในภาคผนวก ข.)

2. กำหนดค่าสัมบูรณ์ของผลต่างระหว่างค่าสัดส่วนสองประชากร $|p_1-p_2|$ มีความแตกต่างกันเท่ากับ .1 ถึง .8 โดยค่าเพิ่มทีละ .1 และ p_1, p_2 มีค่าตั้งแต่ .1 ถึง .9 โดยค่าเพิ่มทีละ .1 $(p_1 < p_2)$

3. กำหนดค่าสัมประสิทธิ์ความเชื่อมั่น $(1-\alpha)100\%$ เท่ากับ 90%, 95% และ 99%

4. ในการวิจัยครั้งนี้สร้างแบบจำลองข้อมูล โดยใช้เทคนิคมอนติคาร์โลซิมูเลชัน (Monte Carlo Simulation Technique) เขียนโปรแกรมด้วยภาษา Fortran 77 ทำการทดลองซ้ำ 20,000 ครั้ง ในแต่ละสถานการณ์ของการทดลอง

ประโยชน์ที่คาดว่าจะได้รับ

1. ผลที่ได้จากการวิจัยในนี้ จะให้แนวทางในการเลือกใช้วิธีการประมาณแบบช่วงที่เหมาะสมสำหรับผลต่างระหว่างค่าสัดส่วนของสองประชากร

2. เป็นแนวทางในการศึกษาเปรียบเทียบหาวิธีการประมาณค่าแบบช่วงด้วยวิธีอื่นๆ สำหรับผลต่างระหว่างค่าสัดส่วนสองประชากรต่อไป

คำจำกัดความ

ช่วงความเชื่อมั่น (Confidence interval) หมายถึงช่วงค่าประมาณพารามิเตอร์ที่คำนวณได้จากตัวอย่างหนึ่งชุดใดๆ ที่ระดับความเชื่อมั่นที่กำหนด

การประมาณค่าแบบช่วง (Interval estimation) หมายถึงการประมาณที่จะให้ช่วงๆหนึ่งซึ่งมีคุณสมบัติว่าค่าที่แท้จริงของประชากรจะอยู่ในช่วงที่ประมาณได้ด้วยความเชื่อมั่นระดับหนึ่ง โดยที่การประมาณค่าแบบช่วงเป็นการประมาณโดยอาศัยตัวประมาณแบบจุด และการแจกแจงความน่าจะเป็นของตัวประมาณนั้น

สัดส่วนประชากร (Population proportion) หมายถึงอัตราส่วนของจำนวนความสำเร็จทั้งหมดในประชากรกับจำนวนประชากรทั้งหมด ซึ่งโดยปกติมักไม่ทราบค่าที่แท้จริง

สัมประสิทธิ์ความเชื่อมั่น (Confidence coefficient) หมายถึงความน่าจะเป็นที่ช่วงสุ่มจะครอบคลุมค่าของพารามิเตอร์