บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการประมาณค่าพารามิเตอร์ของประชากรนั้นเราสามารถทำการประมาณค่าพารามิเตอร์ได้ 2 แบบ คือ การประมาณค่าแบบจุด(point estimation) และการประมาณค่าแบบช่วง(interval estimation) การประมาณค่าแบบจุด เป็นการประมาณค่าพารามิเตอร์ด้วยค่าประมาณเพียงค่าเดียว ซึ่งเราไม่สามารถระบุได้ว่าค่าประมาณที่ได้จากตัวอย่างนั้นเท่ากันกับค่าจริงของพารามิเตอร์จากประชากร เพียงแต่คาดเดาได้ว่าค่าประมาณที่ได้ไม่น่าจะห่างไกลจากค่าจริงเท่านั้น ซึ่งขึ้นอยู่กับการใช้ตัวประมาณที่เหมาะสมซึ่งมีหลักเกณฑ์ในการคัดเลือกตัวประมาณหลายประการ เช่น ความพอเพียง ความไม่เอนเอียง ความคงเส้นคงวา และความมีประสิทธิภาพเป็นต้น อย่างไรก็ตามการประมาณค่าแบบจุดเราไม่สามารถระบุระดับความน่าจะเป็นของความผิดพลาดได้(probability of error) ซึ่งถ้าเราต้องการการประมาณค่าแบบช่วงจะสามารถระบุระดับความน่าจะเป็นของความผิดพลาดได้ การประมาณแบบช่วงเป็นการประมาณที่จะให้ช่วงหนึ่งซึ่งมีคุณสมบัติว่า ช่วงค่าประมาณที่ได้สามารถคลุมค่าที่แท้จริงของพารามิเตอร์ได้ด้วยค่าความเชื่อมั่นระดับหนึ่ง โดยที่การประมาณค่าแบบช่วงเป็นการประมาณโดยอาศัยตัวประมาณแบบจุด และการแจกแจงของตัวประมาณนั้น ซึ่งผลของการประมาณจะทำให้ผู้วิจัยเชื่อมั่นว่าช่วงที่ประมาณได้จะครอบคลุมค่าพารามิเตอร์ของประชากรที่เราสนใจ ผลกระทบต่อความกว้างของช่วงความเชื่อมั่นคือ ระดับความเชื่อมั่นที่ใช้ในการคำนวณ การแจกแจงของตัวสถิติความสอดคล้องกับข้อสมมติ(assumption) และขนาดตัวอย่างเป็นต้น ดังนั้นจึงควรเลือกใช้ช่วงความเชื่อมั่นที่มีความน่าเชื่อถือมากที่สุดในสถานการณ์ต่างๆ

การแจกแจงแบร์นูลลี(Bernoulli distribution) เป็นการแจกแจงที่อธิบายถึงเหตุการณ์ของการทดลองใดๆ ที่มีผลการทดลองที่เป็นไปได้แค่ 2 ทาง คือ ลักษณะที่เราสนใจ และ ลักษณะที่เราไม่สนใจ โดยเกิดลักษณะที่สนใจ ด้วยความน่าจะเป็น p และ เกิดลักษณะที่ไม่สนใจ ด้วยความน่าจะเป็น $1-p$ และเมื่อทำการทดลองแบร์นูลลี(Bernoulli trial) n ครั้ง โดยที่การทดลองแต่ละครั้งอิสระต่อกันและมีความน่าจะเป็นที่จะเกิดลักษณะที่สนใจเท่ากับ p คงที่ตลอดการทดลอง จะได้ว่าจำนวนครั้งที่เกิดลักษณะที่เราสนใจ ที่เกิดขึ้นจากการทำการทดลอง n ครั้ง จะมีการแจกแจงทวินาม(Binomial distribution) ที่มีพารามิเตอร์คือ n, p

ข้อมูลแบบจับคู่(paired data) ของการแจกแจงแบร์นูลลีเป็นลักษณะของข้อมูลที่เป็นค่าสังเกตที่เป็นไปได้แค่ 2 ทางคือ ลักษณะที่เราสนใจ และ ลักษณะที่เราไม่สนใจ ที่มาจากหน่วยตัวอย่างเดียวกัน ฝาแฝดสามีภรรยา พี่น้อง เช่น ประสิทธิภาพการใช้งานคอมพิวเตอร์ก่อนและหลังการอบรมการของนักศึกษา ข้อมูลที่รวบรวมได้นี้จะมีค่า เป็น 0 และ 1 โดยที่ 0 จะหมายถึง การใช้งานคอมพิวเตอร์ของนักศึกษาไม่ผ่านเกณฑ์ 1 จะหมายถึง การใช้งานคอมพิวเตอร์ของนักศึกษาที่ผ่านเกณฑ์ จะสังเกตเห็นว่าประชากรเป็นประชากรที่ไม่เป็นอิสระต่อกัน ลักษณะข้อมูลเป็นดังนี้

เก็บข้อมูลจากประชากรที่มีการแจกแจงแบร์นูลลีมาเป็นคู่ n คู่

$$(x_{01}, x_{11}), (x_{02}, x_{12}), \ldots, (x_{0n}, x_{1n})$$

โดยที่

x_{0j} หมายถึง ข้อมูลที่สุ่มมาจากประชากรที่ 1 หน่วยตัวอย่างที่ j

x_{1j} หมายถึง ข้อมูลที่สุ่มมาจากประชากรที่ 2 หน่วยตัวอย่างที่ j

สำหรับประชากรที่ 1 กำหนดให้ $X_{01}, X_{02}, \ldots, X_{0n}$ เป็นตัวแปรสุ่มที่มีการแจกแจงแบร์นูลลี ที่เป็นอิสระกันโดยมีค่าพารามิเตอร์ p_0 และประชากรที่ 2 กำหนดให้ $X_{11}, X_{12}, \ldots, X_{1n}$ เป็นตัวแปรสุ่มที่มีการแจกแจงแบร์นูลลี ที่เป็นอิสระกันโดยมีค่าพารามิเตอร์ p_1 โดยที่ $0 \le p_0, p_1 \le 1$ จะได้ว่า $Y_0 = \sum_{j=1}^{n} X_{0j}$ และ $Y_1 = \sum_{j=1}^{n} X_{1j}$ เป็นจำนวนครั้งทั้งหมดของการเกิดลักษณะที่สนใจจากการทดลอง n ครั้ง ดังนั้น Y_0 และ Y_1 มีการแจกแจงทวินามที่มีค่าพารามิเตอร์ n, p_0 และ n, p_1 ตามลำดับ เมื่อ $n \to \infty$ จะได้ว่า Y_0 และ Y_1 จะมีการแจกแจงแบบปกติโดยประมาณที่มีค่าเฉลี่ย np_0 และความแปรปรวน $np_0(1-p_0)$ และมีค่าเฉลี่ย np_1 และความแปรปรวน $np_1(1-p_1)$ ตามลำดับ

ในการประมาณค่าพารามิเตอร์ p_0, p_1 หรือค่าสัดส่วนของประชากร(population proportion) ของประชากรที่ 1 และ 2 ตามลำดับ ในการประมาณค่าแบบจุดจะได้ว่า $\hat{p}_0 = \frac{Y_0}{n}$ เป็นตัวประมาณภาวะน่าจะเป็นสูงสุด(maximum likelihood estimator) และไม่เอนเอียง (Unbiased Estimator) ของพารามิเตอร์ p_0 และ $\hat{p}_1 = \frac{Y_1}{n}$ เป็นตัวประมาณภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ p_1

จากทฤษฎีบทลิมิตเข้าสู่ส่วนกลาง(the central limit theorem) จะได้ว่า $\hat{p}_0, \hat{p}_1$ มีการแจกแจงปกติโดยประมาณที่มี ค่าเฉลี่ย p_0 และความแปรปรวน $\frac{p_0(1-p_0)}{n}$ และ มีค่าเฉลี่ย p_1 และความแปรปรวน $\frac{p_1(1-p_1)}{n}$ ตามลำดับ สำหรับตัวอย่างสุ่มทั้ง 2 ชุดที่มีความอิสระต่อกันและขนาดตัวอย่างเท่ากัน n ดังนั้นเมื่อ $n \to \infty$ จากทฤษฎีบทลิมิตเข้าสู่ส่วนกลาง จะได้ว่าการแจก

แจงของผลต่างระหว่างค่าสัดส่วนของสองตัวอย่าง $\hat{p}_1 - \hat{p}_0$ มีการแจกแจงปกติโดยประมาณที่มีค่าเฉลี่ย $p_1 - p_0$ และความแปรปรวน $\frac{p_1(1-p_1)+p_0(1-p_0)}{n}$ และเมื่อกำหนดระดับนัยสำคัญ α สามารถหาช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับผลต่างค่าสัดส่วนแบร์นูลลีที่เป็นอิสระต่อกัน $p_1 - p_0$ ได้คือ

$$\left[\hat{p}_1 - \hat{p}_0 - z_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \sqrt{\hat{p}_1(1-\hat{p}_1)+\hat{p}_0(1-\hat{p}_0)},\right.$$
$$\left.\hat{p}_1 - \hat{p}_0 + z_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \sqrt{\hat{p}_1(1-\hat{p}_1)+\hat{p}_0(1-\hat{p}_0)}\right]$$

เมื่อ Y_0 และ Y_1 มาจากการแจกแจงแบร์นูลลีแบบจับคู่ จะได้ $\hat{p}_0$ และ $\hat{p}_1$ ไม่อิสระกัน ดังนั้น $Var(\hat{p}_1 - \hat{p}_0) = Var(\hat{p}_1) + Var(\hat{p}_1) - 2Cov(\hat{p}_0, \hat{p}_1)$ โดยที่ $Cov(\hat{p}_0, \hat{p}_1) \neq 0$ ดังนั้นไม่สามารถใช้ช่วงความเชื่อมั่นของค่าสัดส่วนแบร์นูลลีที่เป็นอิสระต่อกัน

การที่ตัวอย่างสุ่มทั้ง 2 ชุดมีความสัมพันธ์แบบจับคู่ ดังนั้นเมื่อ $n \to \infty$ จากทฤษฎีบทลิมิตเข้าสู่ส่วนกลาง จะได้ว่าการแจกแจงของผลต่างระหว่างค่าสัดส่วนของสองตัวอย่าง $\hat{p}_1 - \hat{p}_0$ มีการแจกแจงปกติโดยประมาณที่มี ค่าเฉลี่ย $p_1 - p_0$ และความแปรปรวน $\frac{\hat{p}_0(1-\hat{p}_0)+\hat{p}_1(1-\hat{p}_1)+2(\hat{p}_0\hat{p}_1 - \hat{p}_{11})}{n}$ และเมื่อกำหนดระดับนัยสำคัญ α สามารถหาช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ $p_1 - p_0$ ได้คือ

$$\left[\hat{p}_1 - \hat{p}_0 - z_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \sqrt{\hat{p}_0(1-\hat{p}_0)+\hat{p}_1(1-\hat{p}_1)+2(\hat{p}_0\hat{p}_1 - \hat{p}_{11})},\right.$$
$$\left.\hat{p}_1 - \hat{p}_0 + z_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \sqrt{\hat{p}_0(1-\hat{p}_0)+\hat{p}_1(1-\hat{p}_1)+2(\hat{p}_0\hat{p}_1 - \hat{p}_{11})}\right]$$

โดยที่ $\hat{p}_{11} = \frac{Y_{11}}{n}$, $Y_{11} = \sum_{j=1}^{n} X_{0j} X_{1j}$

และ $z_{1-\frac{\alpha}{2}}$ คือเปอร์เซ็นต์ไทล์(percentile)ที่ $\left(1-\frac{\alpha}{2}\right)100\%$ ของการแจกแจงปกติมาตรฐานที่มีค่าเฉลี่ย 0 และความแปรปรวนเป็น 1 ช่วงที่คำนวณได้นี้เรียกว่า ช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ รูปแบบวิธีการประมาณข้างต้นจะเรียกว่า Wald interval (WA) สามารถให้ช่วงการประมาณที่ดีหรือให้ค่าสัมประสิทธิ์ความเชื่อมั่นไม่ต่ำกว่า

ค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด เมื่อ $n \to \infty$ วิธีนี้เป็นวิธีที่นิยมใช้กันเพราะวิธีนี้ เป็นวิธีที่มีพื้นฐานมาจาก ทฤษฎีอะซิมโตติค(asymptotic theory) อย่างไรก็ตามในทางปฏิบัติการรวบรวมข้อมูลจากตัวอย่างอาจไม่สามารถสุ่มตัวอย่างที่มีขนาดใหญ่ได้ เนื่องจากอาจจะมีข้อจำกัดหลายอย่างอาทิเช่น งบประมาณ เวลา หน่วยตัวอย่างที่จำกัด เป็นต้น ดังนั้นการประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนทวินามของข้อมูลแบบจับคู่ โดยถ้าใช้ช่วงความเชื่อมั่น Wald's method ในขณะที่ n มีขนาดเล็ก อาจจะให้ช่วงความเชื่อมั่นที่ได้จะไม่มีคุณสมบัติตามที่ต้องการ

นักสถิติหลายคนได้เสนอวิธีการต่างๆในการประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ เมื่อขนาดตัวอย่างเล็ก เช่น

Newcombe (1998) ได้พิจารณาช่วงความเชื่อมั่นที่มีประโยชน์หลายๆวิธีทั้งหมด10 วิธี เช่น Wald interval, Wald interval with continuity, Score interval เป็นต้น และหลังจากสรุปผลจากการศึกษา Newcombeได้แนะนำ score interval with continuity(Newcombe's method)ซึ่งมีพื้นฐานมากจากวิธีวิลสันสกอร์ (Wilson score method) สำหรับสัดส่วนของประชากรเดียว

May and Johnson (1997) ได้สร้างความเชื่อมั่นที่น่าสนใจขึ้นมาโดยสร้างมาจากส่วนกลับของตัวสถิติที่ใช้สำหรับการทดสอบสมมติฐาน (inverting a test statistics)

Zhou and Qin (2003) ได้สร้างช่วงความเชื่อมั่นใหม่สำหรับความแตกต่างระหว่างสัดส่วนแบร์นูลลีจับคู่เพื่อแก้ความเบ้ของการแจกแจงของstudentized differenceใน Edgeworth expansionโดยใช้ monotone transformation

ความเป็นมาดังกล่าวข้างต้นผู้วิจัยจึงสนใจการประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่เมื่อขนาดตัวอย่างมีขนาดเล็ก และเปรียบเทียบวิธีการประมาณช่วงความเชื่อมั่นทั้ง 4 วิธีเพื่อหาข้อสรุปในการนำวิธีการประมาณแต่ละวิธีที่มีความเหมาะสมไปใช้ในแต่ละสถานการณ์วิธีการ 4 วิธีคือ

1. วิธีการประมาณของ Wald (WA)
2. วิธีการประมาณของ Newcombe (NH)
3. วิธีการประมาณของ May และ Johnson (MJ)
4. วิธีการประมาณของ Zhou และ Qin (ZQ)

1.2 วัตถุประสงค์การวิจัย

วัตถุประสงค์ของการวิจัยมีดังนี้

1. เพื่อศึกษาและเปรียบเทียบการประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ว่าโดยพิจารณาสัมประสิทธิ์ช่วงความเชื่อมั่น และความยาวเฉลี่ยน้อยที่สุดของช่วงความเชื่อมั่น ของ 4 วิธีคือ
    1. วิธีการประมาณของ Wald (WA)
    2. วิธีการประมาณของ Newcombe (NH)
    3. วิธีการประมาณของ May และ Johnson (MJ)
    4. วิธีการประมาณของ Zhou และ Qin (ZQ)
2. เพื่อทราบปัจจัยที่มีผลต่อช่วงความเชื่อมั่นของแต่ละวิธี
3. เพื่อหาข้อเสนอแนะในการใช้ช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่

1.3 สมมติฐานการวิจัย

การประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่โดยใช้วิธีการประมาณของ Zhou และ Qin (ZQ) จะให้ค่าสัมประสิทธิ์ความเชื่อมั่นไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด และให้ความยาวเฉลี่ยของช่วงความเชื่อมั่นต่ำกว่าวิธีการประมาณอื่นๆ

1.4 ข้อตกลงเบื้องต้น

การวิจัยครั้งนี้มีข้อตกลงเบื้องต้นสำหรับการดำเนินการวิจัย ดังนี้

1. ในแต่ละประชากรมีการแจกแจงแบร์นูลลีที่มีความสัมพันธ์กันแบบจับคู่
2. ในแต่ละประชากรจะมีการสุ่มขนาดตัวอย่าง n เท่ากัน
3. ค่าสัดส่วนของความสำเร็จในแต่ละประชากร (p_0, p_1) เป็นพารามิเตอร์ไม่ทราบค่า

1.5 ขอบเขตของการวิจัย

ขอบเขตเบื้องต้นของการศึกษาการประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ ซึ่งประชากรทั้ง 2 ที่สุ่มมาจะมีความสัมพันธ์กัน ให้ตัวอย่างของประชากรกลุ่มที่ 1 และ ประชากรกลุ่มที่ 2 มีขนาด n เท่ากันและความน่าจะเป็นที่เกิดลักษณะที่เราสนใจของประชากรกลุ่มที่ 1 และ ประชากรกลุ่มที่ 2 เท่ากับ p_0, p_1 ตามลำดับ

กำหนดค่าพารามิเตอร์ต่างๆเบื้องต้นดังนี้

1. กำหนดให้ขนาดตัวอย่างที่ใช้ $n = n_1 = n_2$ มีค่าตั้งแต่ 10,20,30,40,50,60,70,80

2. กำหนดค่าความน่าจะเป็นร่วมที่ทำให้ได้ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างประชากรกลุ่มที่ 1 และ 2 มีค่าอยู่ใน ระดับต่ำ $0 \le r < 0.4$, ระดับกลาง $0.4 \le r \le 0.6$, ระดับสูง $0.6 < r \le 1$

3. กำหนดระดับความเชื่อมั่น 3 ระดับคือ 90%, 95% และ 99%

4. กำหนดค่า p_0, p_1 ให้มีค่าตั้งแต่ 0.1 ถึง 0.9 โดยเพิ่มค่าทีละ 0.1 ซึ่งจะให้ค่าผลต่างระหว่างค่าสัดส่วนของสองประชากร $(p_1 - p_0)$ มีความแตกต่างกันตั้งแต่ 0 ถึง 0.8 โดยเพิ่มค่าทีละ 0.1 และกำหนดให้ $p_1 \ge p_0$

5. ข้อมูลที่ใช้ในการวิจัยนี้จะจำลองข้อมูลโดยใช้เทคนิคการจำลองแบบมอนติคาร์โล

1.6 เกณฑ์ที่ใช้ในการประเมินประสิทธิภาพของช่วงความเชื่อมั่น

หลักเกณฑ์ในการพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ มีดังนี้

1. เปรียบเทียบค่าสัมประสิทธิ์ความเชื่อมั่น(confidence coefficient) ที่คำนวณได้จากแต่ละวิธีการประมาณกับค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด ในการตรวจสอบว่าวิธีการใดที่ให้ค่าสัมประสิทธิ์ความเชื่อมั่นจากการทดลองไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดได้หรือไม่นั้น ผู้วิจัยอาศัยการทดสอบสมมติฐานโดยใช้ตัวสถิติ Z ที่ระดับนัยสำคัญ 0.05 (รายละเอียดอยู่ในหัวข้อที่ 2.10)

2. ค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นที่ต่ำกว่าเป็นวิธีการที่มีประสิทธิภาพมากกว่า ทั้งนี้ในการเปรียบเทียบค่าความยาวเฉลี่ยของช่วงความเชื่อมั่น จะเปรียบเทียบเฉพาะในกรณี

วิธีการประมาณที่ให้ค่าสัมประสิทธิ์ความเชื่อมั่นจากการทดลองไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด

1.7 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับของการวิจัยในครั้งนี้

1. เพื่อสามารถประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่แบบช่วงทั้ง 4 วิธีได้

2. ผลที่ได้จากการวิจัยในครั้งนี้จะเป็นแนวทางในการตัดสินใจเลือกใช้วิธีประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่ในแต่ละสถานการณ์ต่างๆ

3. เป็นแนวทางในการศึกษาเปรียบเทียบวิธีการประมาณช่วงความเชื่อมั่นอื่นๆสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่