บทที่ 2

ทฤษฎีที่เกี่ยวข้องกับการวิจัย

2.1 การแจกแจงแบร์นูลลี

การทดลองใดๆที่ผลการทดลองนั้นเป็นไปได้ 2 ลักษณะคือ ลักษณะที่เราสนใจและลักษณะที่เราไม่สนใจเท่านั้น เราจะเรียกการทดลองนั้นว่า การทดลองแบร์นูลลี เช่นการทดลองโยนเหรียญ 1 อัน 1 ครั้ง ผลการทดลองที่ได้จะเป็น 2 ลักษณะคือ หัว และ ก้อย อย่างใดอย่างหนึ่ง ดังนั้นถ้าลักษณะที่เราสนใจคือ การที่เหรียญขึ้น และ การที่เหรียญขึ้นก้อย คือลักษณะที่เราไม่สนใจ การทดลองนี้ก็จะเป็นการทดลองแบร์นูลลี เป็นต้น

ในการทดลองแบร์นูลลี ถ้าให้ความน่าจะเป็นลักษณะที่เราสนใจ มีค่าเท่ากับ p ดังนั้นความน่าจะเป็นลักษณะที่เราไม่สนใจ มีค่าเท่ากับ $1-p$

กำหนดให้ X เป็นตัวแปรสุ่มของการแจกแจงแบร์นูลลี เขียนแทนด้วย $X \sim Ber(p)$

$$X = \begin{cases} 1 & \text{เมื่อเกิดลักษณะที่สนใจ ด้วยความน่าจะเป็น } p \\ 0 & \text{เมื่อเกิดลักษณะที่ไม่สนใจ ด้วยความน่าจะเป็น } 1-p \end{cases}$$

ฟังก์ชันความน่าจะเป็นของ X คือ $f(x) = p^x(1-p)^{1-x}$, $x = 0,1$

ค่าคาดหวังของ X คือ $E(X) = p$

ค่าความแปรปรวนของ X คือ $Var(X) = p(1-p)$

2.2 การแจกแจงทวินาม

เมื่อเราทำการทดลองแบร์นูลลี n ครั้ง เราสนใจจำนวนครั้งที่เกิดเหตุการณ์ที่เราสนใจโดยไม่คำนึงถึงตำแหน่งของการเกิดเหตุการณ์แต่ละครั้งเป็นอิสระต่อกัน ถ้าจำนวนครั้งที่เกิดเหตุการณ์ที่เราสนใจคือตัวแปรสุ่ม Y ของการทดลองแบร์นูลลี n ครั้ง $(Y = 1,2,...,n)$ ดังนั้น Y เป็นตัวแปรสุ่มของการแจกแจงทวินาม $Y \sim B(n,p)$

ถ้าจากการทดลองเกิดลักษณะที่สนใจ y ครั้ง $(y=1,2,...,n)$และเกิดลักษณะที่ไม่สนใจ $n-y$ ครั้ง ดังนั้นความน่าจะเป็นที่เกิดลักษณะที่สนใจ y ครั้ง คือ $p^y(1-p)^{n-y}$ จากนั้นพิจารณาจำนวนหนทางของการเกิดของลักษณะที่สนใจเป็นไปได้ทั้งหมด $\binom{n}{y}=\frac{n!}{y!(n-y)!}$ หนทาง

ฟังก์ชันความน่าจะเป็นของ Y คือ $f(y)=\binom{n}{y}p^y(1-p)^{n-y}$, $y=0,1,...n$

ค่าคาดหวังของ Y คือ $E(Y)=np$

ค่าความแปรปรวนของ Y คือ $Var(Y)=np(1-p)$

2.3 ความน่าจะเป็นที่มีเงื่อนไข (Conditional Probability)

ให้ A และ B เป็นเหตุการณ์ซึ่ง $P(A)\geq 0$ เราเรียก $P(B/A)$ ว่าเป็น ความน่าจะเป็นที่มีเงื่อนไขของ B เมื่อกำหนด A (the conditional probability of b given a) ถ้า

$$P(B/A)=\frac{P(B\cap A)}{P(A)}, P(A)>0$$

2.4 ความแปรปรวนร่วมและสหสัมพันธ์ (Covariance and Correlation)

ให้ X และ Y เป็นตัวแปรสุ่ม เราเรียกว่า $\sigma_{xy}(\text{cov}(X,Y))$ ว่าเป็น ความแปรปรวนร่วมของ X และ Y (covariance of X and Y) ถ้า

$$\text{cov}(X,Y)=E(XY)-E(X)E(Y)$$

ให้ X และ Y เป็นตัวแปรสุ่ม เราเรียกว่า ρ_{xy} ว่าเป็น สัมประสิทธิ์สหสัมพันธ์ ของ X และ Y (correlation coefficient of X and Y) ถ้า

$$\rho_{xy}=\frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

2.5 การแจกแจงแบร์นูลลีของข้อมูลแบบจับคู่

ให้ X_0 , X_1 มีการแจกแจงร่วมโดยมีฟังก์ชันความน่าจะเป็นร่วม $p_{X_0X_1}(x_0, x_1)$ แสดงเป็นตารางดังนี้

		X_1		
		0	1	$p_{X_0}(x_0)$
X_0	0	p_{00}	p_{01}	$p_{00}+p_{01}=1-p_0$
	1	p_{10}	p_{11}	$p_{10}+p_{11}=p_0$
	$p_{X_1}(x_1)$	$p_{00}+p_{10}=1-p_1$	$p_{01}+p_{11}=p_1$	1

จากการแจกแจงร่วมได้การแจกแจงขอบของ $X_o \sim Ber(p_0)$ และ การแจกแจงขอบของ $X_1 \sim Ber(p_1)$

โดยที่

$p_{11} = p_{X_0,X_1}(X_0 = 1, X_1 = 1)$, $p_{00} = p_{X_0,X_1}(X_0 = 0, X_1 = 0)$

$p_{10} = p_{X_0,X_1}(X_0 = 1, X_1 = 0)$, $p_{01} = p_{X_0,X_1}(X_0 = 0, X_1 = 1)$

$E(X_0) = p_0$, $E(X_1) = p_1$

$Var(X_0) = p_0(1-p_0)$, $Var(X_1) = p_1(1-p_1)$

$E(X_1 - X_0) = p_1 - p_0$, $Cov(X_0, X_1) = p_{11} - p_0p_1$

$Var(X_1 - X_0) = p_1(1-p_1) + p_0(1-p_0) + 2(p_0p_1 - p_{11})$

2.6 การประมาณค่าพารามิเตอร์ของการแจกแจงแบร์นูลลีของข้อมูลแบบจับคู่

การทดลองใดๆที่ผลการทดลองนั้นเป็นไปได้ 2 ลักษณะคือ ลักษณะที่เราสนใจและลักษณะที่เราไม่สนใจ ด้วยความน่าจะเป็น p และ ความน่าจะเป็น $1-p$ ตามลำดับ เราจะเรียกการแจกแจงการทดลองนี้ว่า การแจกแจงแบร์นูลลี $X \sim Ber(p)$, p คือค่าสัดส่วนของประชาการที่เกิดผลสำเร็จ

การประมาณค่าพารามิเตอร์ด้วยความควรจะเป็นสูงสุด (Maximum Likelihood Estimation)

ผู้ค้นพบ วิธีนี้เป็นคนแรกชื่อ Gauss C.F. (1821) ซึ่งเป็นนักคณิตศาสตร์ชาวเยอรมัน ต่อมานักสถิติชาวอังกฤษชื่อ R.A. Fisher (1922) ได้ปรับปรุงวิธีการและตรวจสอบคุณสมบัติต่างๆ วิธีการนี้จะใช้ได้เมื่อ ตัวอย่างสุ่มมีการแจกแจงแบบมีพารามิเตอร์

สมมติให้ $\underset{\sim}{x}$ มีฟังก์ชันความหนาแน่น $L(\theta;\underset{\sim}{x})$ สำหรับ $\underset{\sim}{x}$ ซึ่งคงที่ ดังนั้น $L(\theta;\underset{\sim}{x})$เป็นฟังก์ชันภาวะน่าาจะเป็นซึ่งวัดว่า θ จะมีความควรจะเป็นเท่าใดเมื่อสุ่มตัวอย่าง $\underset{\sim}{x}$ หลักการของวิธีนี้คือพยายามหา $\hat{\theta} = \hat{\theta}(\underset{\sim}{x})$ ซึ่งทำให้ $L(\theta;\underset{\sim}{x})$ มีความควรจะเป็นสูงสุดกล่าวคือ $\hat{\theta}$ สอดค้อลงกับ

$$L(\hat{\theta}(\underset{\sim}{x}),\underset{\sim}{x}) = \max_{\theta} L(\theta;\underset{\sim}{x})$$

และเราเรียก $\hat{\theta}$ ว่าเป็น ตัวประมาณความควรจะเป็นสูงสุด

สำหรับพารามิเตอร์ p เราจะหาตัวสถิติสำหรับ p ด้วยความควรจะเป็นสูงสุดดังนี้

กำหนดให้ $X_1, X_2, ..., X_n$ มีการแจกแจง $Ber(p)$ และอิสระต่อกัน โดยมีฟังก์ชันความน่าจะเป็นในรูป $f(x;p) = p^x(1-p)^{1-x}$, $x = 0,1$ และ $0 \le p \le 1$
ได้ฟังก์ชันความควรจะเป็นคือ

$$\begin{aligned}
L(p;\underset{\sim}{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{x_i} \\
&= p^{\sum x_i}(1-p)^{n-\sum x_i} \\
\ln L(p;\underset{\sim}{x}) &= \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p) \\
\frac{\partial \ln L(p;\underset{\sim}{x})}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\
\frac{\sum_{i=1}^{n} x_i}{p} &= \frac{n - \sum_{i=1}^{n} x_i}{1-p}
\end{aligned}$$

$$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i = np - p\sum_{i=1}^{n} x_i$$

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

ดังนั้น $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$ เป็น ตัวประมาณความควรจะเป็นสูงสุด(Maximum Likelihood Estimator) ของ p

กำหนดให้ $\{X_{01}, X_{02}, ..., X_{0n}\}$ เป็นตัวอย่างขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงแบร์นูลลีที่มีค่าความน่าจะเป็นในการเกิดผลสำเร็จหรือค่าสัดส่วนประชากรเท่ากับ p_0 และ $\{X_{11}, X_{12}, ..., X_{1n}\}$ เป็นตัวอย่างขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงแบร์นูลลี ที่มีค่าความน่าจะเป็นในการเกิดผลสำเร็จหรือค่าสัดส่วนประชากรเท่ากับ p_1 โดยที่ $\{X_{01}, X_{02}, ..., X_{0n}\}$ และ $\{X_{11}, X_{12}, ..., X_{1n}\}$ มีความสัมพันธ์แบบจับคู่

ให้ $Y_0 = \sum_{j=1}^{n} X_{0j}$ ดังนั้น $Y_0 \sim B(n, p_0)$ โดยมีค่าเฉลี่ยเท่ากับ np_0 และ ความแปรปรวนเท่ากับ $np_0(1-p_0)$

ให้ $Y_1 = \sum_{j=1}^{n} X_{1j}$ ดังนั้น $Y_1 \sim B(n, p_1)$ โดยมีค่าเฉลี่ยเท่ากับ np_1 และ ความแปรปรวนเท่ากับ $np_1(1-p_1)$

เมื่อพิจารณาค่าเฉลี่ยของ $\hat{p}_0, \hat{p}_1$ โดยที่ $\hat{p}_0 = \frac{Y_0}{n}$ และ $\hat{p}_1 = \frac{Y_1}{n}$

$$E(\hat{p}_i) = E\left(\frac{Y_i}{n}\right) = \frac{1}{n.} E(Y_i) = \frac{np_i}{n} = p_i \qquad ,i = 0,1$$

แสดงให้เห็นว่า $\hat{p}_0, \hat{p}_1$ เป็นตัวประมาณที่ไม่เอนเอียง(Unbiased Estimator) ของ p_0, p_1 และค่าความแปรปรวนของ $\hat{p}_0, \hat{p}_1$ คือ

$$Var(\hat{p}_i) = Var\left(\frac{Y_i}{n}\right) = \frac{1}{n^2.} Var(Y_i) = \frac{np_i(1-p_i)}{n^2} = \frac{p_i(1-p_i)}{n} \qquad ,i = 0,1$$

เนื่องจาก $\{X_{01}, X_{02}, ..., X_{0n}\}$ และ $\{X_{11}, X_{12}, ..., X_{1n}\}$ เป็นกลุ่มตัวอย่าง 2 กลุ่มที่สุ่มมาจากประชากรที่มีการแจกแจงแบร์นูลลีที่มีความสัมพันธ์กันแบบจับคู่ n คู่ ดังนั้น $Y_0 = \sum_{j=1}^{n} X_{0j}$ และ $Y_1 = \sum_{j=1}^{n} X_{1j}$ เป็นตัวแปรสุ่มที่มีการแจกแจงแบบ $Y_0 \sim B(n, p_0)$ และ $Y_1 \sim B(n, p_1)$ ตามลำดับ ที่มีความสัมพันธ์กัน เมื่อ $\hat{p}_0 = \frac{Y_0}{n}$, $\hat{p}_1 = \frac{Y_1}{n}$ จะได้ว่า

$$E(\hat{p}_1 - \hat{p}_0) = p_1 - p_0$$

$$Var(\hat{p}_1 - \hat{p}_0) = \frac{p_0(1-p_0) + p_1(1-p_1) + 2(p_0 p_1 - p_{11})}{n}$$

โดยที่ $p_{11} = p_{X_0,X_1}(X_0 = 1, X_1 = 1)$

<u>การประมาณค่าพารามิเตอร์ของการแจกแจงแบร์นูลลีของข้อมูลแบบจับคู่</u>

เมื่อเราสุ่มตัวอย่างแล้วข้อมูลที่ได้จะมีลักษณะเป็นคู่ๆ ทั้งหมด n คู่ เราจะนำผลการทดลองมาใส่ในตารางการจรขนาด 2×2 ดังนี้

		การทดลอง 2		
		ไม่สนใจ(0)	สนใจ(1)	รวม
การทดลอง 1	ไม่สนใจ(0)	Y_{00}	Y_{01}	$Y_{00} + Y_{01}$
	สนใจ(1)	Y_{10}	Y_{11}	$Y_{10} + Y_{11} = Y_0$
	รวม	$Y_{00} + Y_{10}$	$Y_{01} + Y_{11} = Y_1$	n

โดยที่

X_{ij} คือตัวอย่างสุ่มจากการแจกแจงแบร์นูลลี

$$X_{ij} = \begin{cases} 1 & \text{เมื่อเกิดลักษณะที่สนใจ ด้วยความน่าจะเป็น } p \\ 0 & \text{เมื่อเกิดลักษณะที่ไม่สนใจ ด้วยความน่าจะเป็น } 1-p \end{cases}, \quad i = 0,1 \quad j = 1,2,\ldots,n$$

$$Y_{11} = \sum_{j=1}^{n} X_{0j} X_{1j}, \qquad Y_{00} = \sum_{j=1}^{n} (1 - X_{0j})(1 - X_{1j})$$

$$Y_{10} = \sum_{j=1}^{n} X_{0j}(1 - X_{1j}), \qquad Y_{01} = \sum_{j=1}^{n} (1 - X_{0j}) X_{1j}$$

$$Y_{00} + Y_{11} + Y_{01} + Y_{10} = n$$

ค่าประมาณของ p_0 คือ $\hat{p}_0 = \frac{Y_0}{n}$, ค่าประมาณของ p_1 คือ $\hat{p}_1 = \frac{Y_1}{n}$

ค่าประมาณของ p_{10} คือ $\hat{p}_{10} = \frac{Y_{10}}{n}$, ค่าประมาณของ p_{10} คือ $\hat{p}_{10} = \frac{Y_{10}}{n}$

ค่าประมาณของ p_{00} คือ $\hat{p}_{00} = \frac{Y_{00}}{n}$, ค่าประมาณของ p_{11} คือ $\hat{p}_{11} = \frac{Y_{11}}{n}$

2.7 การประมาณค่าแบบช่วง

การประมาณค่าแบบช่วงหรือที่เรียกว่า ช่วงความเชื่อมั่น(Confidence Intervals) เป็นการประมาณค่าพารามิเตอร์ของประชากรอยู่ในช่วงใดช่วงหนึ่งโดยใช้ข้อมูลตัวอย่าง และช่วงการประมาณค่าจะบอกถึงค่าต่ำสุดและค่าสูงสุดของพารามิเตอร์ที่เป็นไปได้

กำหนดให้ $X_1, X_2, ..., X_n$ เป็นตัวแปรสุ่มจากการแจกแจงซึ่งมี θ เป็นพารามิเตอร์ที่ไม่ทราบค่า ให้ $t_1(X_1, X_2, ..., X_n), t_2(X_1, X_2, ..., X_n)$ เป็นตัวสถิติที่ $t_1 < t_2$ และ $P\{t_1 < \theta < t_2\} = 1-\alpha$ ช่วงสุ่ม(random interval) ของ (t_1, t_2) ที่ได้ก็คือตัวประมาณค่าแบบช่วง(interval estimator) โดยที่ t_1 คือขีดจำกัดความเชื่อมั่นล่าง(lower confidence limit) และ t_2 คือขีดจำกัดความเชื่อมั่นบน(upper confidence limit) สำหรับพารามิเตอร์ θ และ $1-\alpha$ เรียกว่า สัมประสิทธิ์ความเชื่อมั่น(confidence coefficient) หรือระดับความเชื่อมั่น(confidence level) โดยทั่วไปมักนิยมใช้ระดับความเชื่อมั่นเป็น 90%, 95% และ 99% ยกตัวอย่างเช่น $P\{t_1 < \theta < t_2\} = 0.95$ หมายถึงความน่าจะเป็นที่ค่าของ θ จะมีค่าอยู่ในช่วง (t_1, t_2) มีค่าเท่ากับ 0.95 และค่าความน่าจะเป็นที่ θ จะมีค่ามากกว่า t_2 หรือน้อยกว่า t_1 มีค่าเท่ากับ 0.10

นอกจากนี้แล้ว สิ่งที่จำเป็นที่จะควรพิจารณาถึงสำหรับการประมาณค่าแบบช่วงคือ ความยาวเฉลี่ยของช่วงประมาณ ซึ่งคือค่าคาดหวังของผลต่างของขีดจำกัดความเชื่อมั่นบนและขีดจำกัดความเชื่อมั่นล่าง หรือหาได้จาก $E(t_2 - t_1)$

2.8 ทฤษฎีบทลิมิตเข้าสู่ส่วนกลาง

ถ้า $\bar{Y}$ คือค่าเฉลี่ยของตัวอย่างสุ่ม $Y_1, Y_2, ..., Y_n$ จากการแจกแจงที่มีค่าเฉลี่ย μ มีความแปรปรวน $\sigma^2 < \infty$ และ $n \to \infty$ จะได้ว่า

$$\bar{Y} \sim N\left(\mu, \sigma_{\bar{Y}}^2\right) \qquad \text{(โดยประมาณ)}$$

หรือ $$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \qquad \text{(โดยประมาณ)}$$

นั่นคือเมื่อสุ่มตัวอย่างขนาดใหญ่จากประชากรที่มีการแจกแจงใดๆ ก็ตามที่มีความแปรปรวนเป็นค่าจำกัด แล้ว ค่าเฉลี่ยตัวอย่างจะมีการแจกแจงลู่เข้าสู่การแจกแจงแบบปกติ

2.9 วิธีการประมาณค่าแบบช่วงของผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่

1. วิธีการประมาณของ Wald

ในการประมาณค่าแบบช่วงของผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่เมื่อกำหนดความน่าจะเป็นของความผิดพลาด α หรือ สัมประสิทธิ์ช่วงความเชื่อมั่น $1-\alpha$ (confidence coefficient) และเมื่อสุ่มตัวอย่างขนาดใหญ่โดยใช้ทฤษฎีบทลิมิตเข้าสู่ส่วนกลางจะได้ว่า $\hat{p}_1 - \hat{p}_0$ จะมีการแจกแจงลู่เข้าสู่การแจกแจงแบบปกติ โดยมีค่าเฉลี่ย เป็น $p_1 - p_0$ และความแปรปรวนเท่ากับ

$$\frac{p_0(1-p_0)+p_1(1-p_1)+2(p_0p_1-p_{11})}{n}$$

$$Z=\frac{(\hat{p}_1-\hat{p}_0)-(p_1-p_0)}{\sqrt{p_0(1-p_0)+p_1(1-p_1)+2(p_0p_1-p_{11})/n}} \sim N(0,1)$$

ดังนั้น $$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{(\hat{p}_1-\hat{p}_0)-(p_1-p_0)}{\sqrt{p_0(1-p_0)+p_1(1-p_1)+2(p_0p_1-p_{11})/n}} < Z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

เนื่องจากเราไม่ทราบค่าสัดส่วนประชากร p_0, p_1 จึงไม่สามารถหาค่าความแปรปรวนได้ ดังนั้นจึงใช้ค่าสัดส่วนตัวอย่าง $\hat{p}_0, \hat{p}_1$ ซึ่งเป็นค่าประมาณแทน จะได้ช่วงความเชื่อมั่น $(1-\alpha)100\%$ โดยประมาณสำหรับ $\hat{p}=\hat{p}_1-\hat{p}_0$ คือ

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

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

$$\hat{p}=\hat{p}_1-\hat{p}_0$$

$$Y_i=\sum_{j=1}^{n}X_{ij} \qquad ;i=0,1$$

$$\hat{p}_i=\frac{Y_i}{n} \qquad ,\hat{p}=\hat{p}_1-\hat{p}_0$$

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

2. วิธีการประมาณของ Newcombe

Newcombe (1998) ได้ศึกษาวิธีการหาช่วงความเชื่อมั่นทั้ง 10 วิธี และ Newcombeได้แนะนำเสนอให้ใช้ Score interval with countinuity(Newcombe's method) ได้แนวคิดมาจากวิธีการประมาณแบบช่วงด้วยวิธีรากของสมการกำลังสองสำหรับค่าสัดส่วนประชากร(Score interval) และปรับแก้ไขความไม่ต่อเนื่องโดยใช้ค่า phi coefficient ($\hat{\phi}$)

ช่วงความเชื่อมั่น $(1-\alpha)100\%$ โดยประมาณสำหรับ $p_1 - p_0$ คือ $\hat{\phi}$

$$\left[\hat{p} - \left(\delta_1^2 - 2\hat{\phi}\delta_1\varepsilon_2 + \varepsilon_2^2\right)^{\frac{1}{2}}, \hat{p} + \left(\delta_2^2 - 2\hat{\phi}\delta_2\varepsilon_1 + \varepsilon_1^2\right)^{\frac{1}{2}} \right]$$

โดยที่

$$Y_{00} = \sum_{j=1}^{n} (1 - X_{0j})(1 - X_{1j}) \quad , Y_{10} = \sum_{j=1}^{n} X_{0j}(1 - X_{1j})$$

$$Y_{01} = \sum_{j=1}^{n} (1 - X_{0j}) X_{1j}$$

$$D = (Y_{00} + Y_{10})(Y_{01} + Y_{11})(Y_{00} + Y_{01})(Y_{10} + Y_{11})$$

ให้ l_1, u_1 เป็นคำตอบของ x สมการ

$$\left(x - \frac{Y_{11} + Y_{01}}{n}\right)^2 = \left(z_{1-\frac{\alpha}{2}}\right)^2 \cdot \frac{x(1-x)}{n}$$

ให้ l_2, u_2 เป็นคำตอบของ x สมการ

$$\left(x - \frac{Y_{11} + Y_{10}}{n}\right)^2 = \left(z_{1-\frac{\alpha}{2}}\right)^2 \cdot \frac{x(1-x)}{n}$$

$$\delta_1 = \frac{Y_{11} + Y_{01}}{n} - l_1 \quad , \quad \delta_2 = \frac{Y_{11} + Y_{10}}{n} - l_2$$

$$\varepsilon_1 = u_1 - \frac{Y_{11} + Y_{01}}{n} \quad , \quad \varepsilon_2 = u_2 - \frac{Y_{11} + Y_{10}}{n}$$

$$\hat{\phi} = \begin{cases} (Y_{00}Y_{11} - Y_{10}Y_{01})/\sqrt{D}, & Y_{00}Y_{11} - Y_{10}Y_{01} \le 0 \quad , D > 0 \\ \max(Y_{00}Y_{11} - Y_{10}Y_{01} - n/2, 0)/\sqrt{D}, & Y_{00}Y_{11} - Y_{10}Y_{01} > 0 \quad , D > 0 \\ 0, & D = 0 \end{cases}$$

3. **วิธีการประมาณของMay และ Johnson interval**

Warren L. May and William D. Johnson (1997) ได้สร้างช่วงความเชื่อมั่นขึ้นมาจากการทดสอบสมมติฐานของตารางการจรขนาด 2×2 โดยใช้ตัวสถิติ Chi-Square

ช่วงความเชื่อมั่น $(1-\alpha)100\%$ โดยประมาณสำหรับ p_1-p_0 คือ

$$\left[\max\left\{-1,\frac{\left(-B-\left(B^2-4AC\right)^{\frac{1}{2}}\right)}{2A}\right\},\min\left\{1,\frac{\left(-B+\left(B^2-4AC\right)^{\frac{1}{2}}\right)}{2A}\right\}\right]$$

โดยที่

$$A=\left(1+z^2_{1-\frac{\alpha}{2}}/n\right)\quad,\quad B=-2\left(Y_{01}-Y_{10}\right)/n$$

$$C=\left(Y_{01}/n-Y_{10}/n\right)^2-z^2_{\frac{\alpha}{2}}\left(Y_{01}+Y_{10}\right)/n^2$$

4. **วิธีการประมาณของ Zhou และ Qin**

Xiao-Hua Zhou กับ Gengsheng Qin (2003) ได้สร้างช่วงความเชื่อมั่นใหม่ที่สามารถปรับแก้ความเบ้ของการแจกแจงของ studentized differenceใน Edgeworth expansionโดยใช้ monotone transformation

ช่วงความเชื่อมั่น $(1-\alpha)100\%$ โดยประมาณสำหรับ p_1-p_0 คือ

$$\left[\max\left\{-1,\hat{p}-\frac{\hat{\sigma}}{\sqrt{n}}\cdot g^{-1}\left(z_{1-\frac{\alpha}{2}}\right)\right\},\min\left\{1,\hat{p}-\frac{\hat{\sigma}}{\sqrt{n}}\cdot g^{-1}\left(z_{\frac{\alpha}{2}}\right)\right\}\right]$$

โดยที่

$$\hat{d}=\hat{p}_1\left(1-\hat{p}_1\right)\left(1-2\hat{p}_1\right)-\hat{p}_0\left(1-\hat{p}_0\right)\left(1-2\hat{p}_0\right)+6\left(\hat{p}_1-\hat{p}_0\right)\left(\hat{p}_{11}-\hat{p}_0\hat{p}_1\right)$$

$$\hat{\sigma}=\left(\hat{p}_1\left(1-\hat{p}_1\right)+\hat{p}_0\left(1-\hat{p}_0\right)+2\left(\hat{p}_0\hat{p}_1-\hat{p}_{11}\right)\right)^{\frac{1}{2}}$$

$$\hat{a}=\frac{\hat{d}}{\left(6\hat{\sigma}^2\right)}\quad,\quad\hat{b}=\frac{1-2\hat{p}}{2}-\frac{\hat{d}}{\left(6\hat{\sigma}^2\right)}$$

$$g^{-1}(y)=\frac{\sqrt{n}}{\hat{b}\hat{\sigma}}\left[\left(1+3\left(\hat{b}\hat{\sigma}\left(\frac{y}{\sqrt{n}}-\frac{\hat{a}\hat{\sigma}}{n}\right)\right)^{\frac{1}{3}}-1\right]\quad\text{ถ้า}\quad\hat{b}\hat{\sigma}\neq0$$

$$g^{-1}(y)=y-\frac{\hat{a}\hat{\sigma}}{\sqrt{n}}\quad\text{ถ้า}\quad\hat{b}\hat{\sigma}=0$$

2.10 เกณฑ์การเปรียบเทียบ

การประมาณช่วงความเชื่อมั่นสำหรับผลต่างค่าสัดส่วนแบร์นูลลีของข้อมูลแบบจับคู่แบบช่วงทั้ง 4 วิธีนั้น จะทำการเปรียบเทียบค่าระดับความเชื่อมั่นและความยาวเฉลี่ยของค่าประมาณแบบช่วงที่คำนวณได้จากแต่ละสถานการณ์ทดลองในการทำการทดลอง 2,000 ครั้ง การตรวจสอบว่าวิธีการประมาณใดให้ระดับความเชื่อมั่นจากการทดลองไม่ต่ำกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดได้หรือไม่นั้น ผู้วิจัยจะอาศัยการทดสอบสมมติฐานโดยใช้ตัวสถิติ Z ดังนี้

$$H_0 : c \geq c_0$$
$$H_1 : c < c_0$$

ขอบเขตของการยอมรับสมมติฐานคือ

$$-Z_{1-\alpha_0} < \frac{\hat{c} - c_0}{\sqrt{\frac{c_0(1-c_0)}{n}}}$$

$$c_0 - Z_{1-\alpha_0}\sqrt{\frac{c_0(1-c_0)}{n}} < \hat{c} < 1$$

ฉะนั้นจะได้ช่วงที่ยอมรับสมมติฐานหลัก คือ

$$\left(c_0 - Z_{1-\alpha_0}\sqrt{\frac{c_0(1-c_0)}{n}}, 1\right)$$

เมื่อ α_0 คือระดับนัยสำคัญหรือ Type I error กำหนดในการทดสอบในการวิจัยครั้งนี้โดยกำหนดระดับนัยสำคัญของการทดสอบเท่ากับ 0.05

c คือระดับความเชื่อมั่น

$\hat{c}$ คือระดับความเชื่อมั่นที่ได้จากการทดลองหรือความน่าจะเป็นที่วิธีการประมาณนั้นจะคลุมค่า $p_1 - p_0$ ซึ่งหาได้จากจำนวนครั้งที่คลุมค่า $p_1 - p_0$ หารด้วยจำนวนครั้งในการทดลอง (n)

c_0 คือระดับความเชื่อมั่นที่กำหนด(0.90, 0.95 และ 0.99)

n จำนวนครั้งของการทดลอง(2,000)

1. ที่ระดับความเชื่อมั่น 90%

$H_0 : c \geq 0.90$

$H_1 : c < 0.90$

จะได้ว่าวิธีการให้ค่าระดับความเชื่อมั่นไม่ต่ำกว่าระดับความเชื่อมั่นที่กำหนดถ้า $\hat{c}$ มีค่าอยู่ในช่วง

$$\left(0.90 - 1.645\sqrt{\frac{0.90(0.10)}{2000}}, 1\right)$$

$$(0.8890, 1)$$

2. ที่ระดับความเชื่อมั่น 95%

$H_0 : c \geq 0.95$

$H_1 : c < 0.95$

จะได้ว่าวิธีการให้ค่าระดับความเชื่อมั่นไม่ต่ำกว่าระดับความเชื่อมั่นที่กำหนดถ้า $\hat{c}$ มีค่าอยู่ในช่วง

$$\left(0.95 - 1.645\sqrt{\frac{0.95(0.05)}{2000}}, 1\right)$$

$$(0.9420, 1)$$

3. ที่ระดับความเชื่อมั่น 99%

$H_0 : c \geq 0.99$

$H_1 : c < 0.99$

จะได้ว่าวิธีการให้ค่าระดับความเชื่อมั่นไม่ต่ำกว่าระดับความเชื่อมั่นที่กำหนดถ้า $\hat{c}$ มีค่าอยู่ในช่วง

$$\left(0.99 - 1.645\sqrt{\frac{0.99(0.01)}{2000}}, 1\right)$$

$$(0.9863, 1)$$

เมื่อคำนวณค่าระดับช่วงความเชื่อมั่นของช่วงความเชื่อมั่นเฉลี่ยที่ได้ในแต่ละวิธีที่ทำการศึกษาทั้ง 4 วิธีแล้ว นำค่าระดับความเชื่อมั่นของช่วงความเชื่อมั่นเฉลี่ยที่ได้มาเปรียบเทียบกับค่า 0.8890, 0.9420 และ 0.9863 ที่ค่าระดับสัมประสิทธิ์ความเชื่อมั่นที่กำหนด 90%, 95% และ 99%ตามลำดับ ถ้าวิธีการประมาณที่ให้ค่าสัมประสิทธิ์ความเชื่อมั่นที่ได้จากการทดลองมีค่าไม่ต่ำกว่าค่าระดับความเชื่อมั่นที่กำหนดในสถานการณ์นั้นๆ ขั้นต่อไปจะพิจารณาเปรียบเทียบค่าความยาวเฉลี่ยของช่วงความเชื่อมั่นว่าวิธีที่ทำการศึกษาวิธีใดให้ค่าความยาวเฉลี่ยของความเชื่อมั่นน้อยที่สุดและจะถือว่าวิธีการนั้นเหมาะสมที่สุดสำหรับสถานการณ์นั้นๆ ทั้งนี้ในการเปรียบเทียบค่าความยาวเฉลี่ยของความเชื่อมั่น จะเปรียบเทียบเฉพาะในกรณีที่วิธีการนั้นให้ค่าสัมประสิทธิ์ความเชื่อมั่นที่ได้จากการทดลองมีค่าไม่ต่ำกว่าค่าระดับความเชื่อมั่นที่กำหนดเท่านั้น