$$\bar{x} - \hat{R}\bar{y} = \bar{x} - \frac{\bar{x}}{\bar{y}} \cdot \bar{y} = 0$$

$$\Rightarrow \quad s_d^2 \simeq \frac{1}{n-1} \sum_i^n (x_i - \hat{R}y_i)^2$$

และโดยปกติแล้วเราไม่อาจทราบค่าของ $\bar{Y}$ ดังนั้นจึงแทนที่ $\bar{Y}$ ด้วยค่าประมาณของ $\bar{Y}$ คือ $\bar{y}$ ดังนั้น

$$\hat{V}(\hat{R}) \simeq \frac{1}{n\bar{y}^2} \frac{N-n}{N} \cdot \frac{1}{n-1} \sum_i^n (x_i - \hat{R}y_i)^2 \quad \text{โดยที่ } \hat{R} = \bar{x}/\bar{y}$$

สำหรับสูตรดังกล่าวนี้สามารถจัดให้เป็นรูปที่สะดวกต่อการคำนวณ (Computation Formular) ได้ดังนี้

$$\hat{V}(\hat{R}) \simeq \frac{1}{n\bar{y}^2} \frac{N-n}{N} \frac{1}{n-1} \left\{ \sum_i^n x_i^2 - 2R \sum_i^n x_i y_i + \hat{R}^2 \sum_i^n y_i^2 \right\}$$

ตัวอย่าง 2.3 จากการสำรวจจำนวนนักเรียนในโรงเรียนมัธยมศึกษาในเขตกทม. ซึ่งมีอยู่ทั้งสิ้น 468 โรงเรียน (ข้อมูลสมมุติ) การสำรวจโดยใช้วิธี SRS โดยสุ่มตัวอย่างโรงเรียนมาจากกรอบตัวอย่าง 100 โรง ปรากฏว่าเป็นโรงเรียนราษฎร์ 46 โรง โรงเรียนรัฐบาล 54 โรง ผลจากการสำรวจปรากฏดังนี้

	n	จำนวนครู	จำนวนนักเรียน	Σy^2	Σx^2	Σxy
รร.รัฐบาล	54	2,024	31,281	111,090	29,881,219	1,729,349
รร.ราษฎร์	46	1,075	13,707	33,119	6,366,785	431,041

ก. ในโรงเรียนแต่ละประเภทจงกะประมาณอัตราส่วนระหว่าง นักเรียน-ครู

ข. จงกะประมาณอัตราส่วนนักเรียน-ครูในข้อ ก. ด้วยช่วงเชื่อมั่น 95%

ค. จงกะประมาณอัตราส่วนระหว่างนักเรียน-ครู ในเขต กทม. พร้อมทั้งกะประมาณอัตราส่วนดังกล่าวด้วยช่วงเชื่อมั่น 99%

วิธีทำ

ก. ให้ R_1 = อัตราส่วนระหว่างนักเรียน-ครูในโรงเรียนรัฐบาล

$$= \frac{\text{จำนวนนักเรียนในโรงเรียนรัฐบาล}}{\text{จำนวนครูในโรงเรียนรัฐบาล}} = \sum_{i}^{N} x_i / \sum_{i}^{N} y_i$$

R_2 = อัตราส่วนระหว่างนักเรียน-ครูในโรงเรียนราษฎร์

$$= \frac{\text{จำนวนนักเรียนในโรงเรียนราษฎร์}}{\text{จำนวนครูในโรงเรียนราษฎร์}} = \sum_{i}^{N} x_i / \sum_{i}^{N} y_i$$

โดยอาศัยข้อมูลจากกลุ่มตัวอย่างโรงเรียนขนาด n = 100 โรงเรียน จะพบว่า

$$\hat{R}_1 = \sum_{i}^{54} x_i / \sum_{i}^{54} y_i = 31,281/2,024 = 15.455$$

$$\hat{R}_2 = \sum_{i}^{46} x_i / \sum_{i}^{46} y_i = 13,707/1,075 = 12.751$$

จากผลการสำรวจชี้ให้เห็นดังนี้คือ

1. อัตราส่วนระหว่างนักเรียน-ครูในโรงเรียนรัฐบาล คือครูหนึ่งคนต่อนักเรียนประมาณ 15 คนหรือครู 1000 คนต่อนักเรียน 15,455 คน

2. อัตราส่วนระหว่างนักเรียน-ครูในโรงเรียนราษฎร์คือครูหนึ่งคนต่อนักเรียนประมาณ 13 คน หรือครู 1000 คนต่อนักเรียน 12,751 คน

ข. การกะประมาณค่า R_1 และ R_2 ด้วยช่วงเชื่อมั่น

1. ในโรงเรียนรัฐบาล (R_1)

$$\hat{V}(\hat{R}_1) = \frac{1}{n\bar{y}^2}\frac{N-n}{N}\cdot\frac{1}{n-1}\left(\sum_i^n x_i^2 - 2\hat{R}_1\sum_i^n x_iy_i + \hat{R}_1^2\sum_i^n y_i^2\right)$$

$N = 468$, $n = 54$, $\hat{R}_1 = 15.455$, $\hat{R}_1^2 = 238.857$

$\bar{y}$ = อัตราเฉลี่ยจำนวนครูโรงเรียนรัฐบาลในกทม.ต่อ 1 โรงเรียน $= \frac{2024}{54} = 37.48$

$\bar{y}^2 = 1{,}404.86$ $\qquad \Sigma x_i^2 = 29{,}881{,}219$

$\Sigma y_i^2 = 111{,}090$ $\qquad \Sigma x_iy_i = 1{,}729{,}349$

ดังนั้น $\hat{V}(\hat{R}_1) = \frac{1}{(54)(1404.86)}\cdot\frac{468-54}{468}\cdot\frac{1}{53}\{29881219 - 2(15.455)(1729349) + (238.857)(111090)\}$

$= .6516$

$s_{\hat{R}_1} = .8072$

ดังนั้น $13.87 < R_1 < 17.04$

หรือสามารถเชื่อถือได้ถึง 95% ว่าอัตราส่วนจริงระหว่างนักเรียน-ครูของโรงเรียนรัฐบาลในเขตกทม.จะปรากฏอยู่ระหว่างครู 1 คนต่อนักเรียนประมาณ 14 คนถึง 17 คน

2. ในโรงเรียนราษฎร์ (R_2)

$$\hat{V}(\hat{R}_2) = \frac{1}{n\bar{y}^2}\cdot\frac{N-n}{N}\frac{1}{n-1}\left(\sum_i^n x_i^2 - 2R_2\sum_i^n x_iy_i + \hat{R}_2^2\sum_i^n y_i^2\right)$$

$N = 468$, $n = 46$, $\hat{R}_2 = 12.751$ $\hat{R}_2^2 = 162.588$

$\bar{y}$ = อัตราเฉลี่ยจำนวนครูโรงเรียนราษฎร์ในกทม.ต่อ 1 โรงเรียน

$$= \frac{1075}{46} = 23.37 \qquad \bar{y}^2 = 546.137, \qquad \Sigma x_i^2 = 6{,}366{,}785$$

$$\sum_i^n y_i^2 = 33{,}119, \qquad \sum_i^n x_i y_i = 431{,}041$$

$$\hat{V}(\hat{R}_2) = \frac{1}{(46)(546.137)} \cdot \frac{468-46.1}{468} \cdot \frac{1}{45} \{6366785 - 2(12.751)(431041) + (162.588)(33119)\}$$

$$s = .7781$$

ดังนั้น $11.22 < R_2 < 14.28$

หรือสามารถเชื่อถือได้ถึง 95% ว่าอัตราส่วนระหว่างนักเรียน-ครูของโรงเรียนราษฎร์ในเขตกทม. จะปรากฏอยู่ในระหว่างครู 1 คนต่อนักเรียนประมาณ 11 ถึง 14 คน

สำหรับข้อ ค. นั้น จะขอเว้นไว้เป็นแบบฝึกหัด การวิเคราะห์และอภิปรายผลสามารถกระทำได้โดยง่ายเช่นเดียวกับข้อ ก. และข้อ ข.

หมายเหตุ สำหรับกรณีของอัตราส่วน W.G. Cochran แนะนำว่า ถ้า $n > 30$ และ C.V. ของทั้ง $\bar{x}$ และ $\bar{y}$ มีค่าไม่เกิน 0.1 กล่าวคือ $\frac{s_x}{\bar{x}} \leq 0.1$ และ $\frac{s_y}{\bar{y}} \leq 0.1$ แล้ว Sampling Distribution ของ $\hat{R}$ จะมีการแจกแจงแบบปกติโดยประมาณคือ $\hat{R} \sim N(R, V(\hat{R}))$ ด้วยเหตุนี้เราจึงสามารถประมาณค่าพารามิเตอร์ R ด้วยช่วงเชื่อมั่นได้ ทั้งยังสามารถทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ R ได้อีกด้วย

กรณีการประมาณค่าด้วยอัตราส่วนนี้ นักศึกษาจะพบว่า $\hat{R}$ เป็นตัวประมาณค่าที่มีอคติของ R หรือ $E(\hat{R}) \neq R$ เว้นแต่เมื่อขนาดตัวอย่าง n ใหญ่มาก เราจึงขออนุโลมว่าให้ใช้ $\hat{R}$ เป็นตัวประมาณค่าของ R ได้

ปริมาณของความเอียงเฉ (Bias) หรือคติของ $\hat{R}$ วัดได้ดังนี้

$$\because \quad \hat{R} = \frac{\bar{x}}{\bar{y}}$$

$$\hat{R} = \frac{\bar{x}}{\bar{y}} = \frac{\bar{X}}{\bar{Y}}\left(1 + \frac{\bar{x}-\bar{X}}{\bar{X}}\right)\left(1 + \frac{\bar{y}-\bar{Y}}{\bar{Y}}\right)^{-1}\ [1]$$

โดยอาศัยการกระจายเทอมของเทเลอร์

$$\left(1 + \frac{\bar{y}-\bar{Y}}{\bar{Y}}\right)^{-1} = 1 - \frac{\bar{y}-\bar{Y}}{\bar{Y}} + \frac{(\bar{y}-\bar{Y})^2}{\bar{Y}^2} - \ldots$$

ดังนั้น $\hat{R}$ $= \frac{\bar{X}}{\bar{Y}}\left(1 + \frac{\bar{x}-\bar{X}}{\bar{X}}\right)\left(1 - \frac{\bar{y}-\bar{Y}}{\bar{Y}} + \frac{(\bar{y}-\bar{Y})^2}{\bar{Y}^2} - \ldots\right)$

$$= R\left(1 + \frac{\bar{x}-\bar{X}}{\bar{X}} - \frac{\bar{y}-\bar{Y}}{\bar{Y}} + \frac{(\bar{y}-\bar{Y})^2}{\bar{Y}^2} - \frac{(\bar{x}-\bar{X})(\bar{y}-\bar{Y})}{\bar{X}\bar{Y}} + \ldots\right)$$

$$\therefore \quad E(\hat{R}) = R + R\cdot\frac{1}{\bar{X}}E(\bar{x}-\bar{X}) - R\cdot\frac{1}{\bar{Y}}E(\bar{y}-\bar{Y}) + R\frac{1}{\bar{Y}^2}E(\bar{y}-\bar{Y})^2$$

$$- R\cdot\frac{1}{\bar{X}\bar{Y}}E(\bar{x}-\bar{X})(\bar{y}-\bar{Y})$$

$$= R + 0 - 0 + \frac{R}{\bar{Y}^2}V(\bar{y}) - \frac{R}{\bar{X}\bar{Y}}E(\bar{x}-\bar{X})(\bar{y}-\bar{Y})$$

$$= R + \frac{R}{\bar{Y}^2}\cdot\frac{N-n}{N}\cdot\frac{S_y^2}{n} - \frac{R}{\bar{X}\bar{Y}}\cdot\frac{N-n}{N}\cdot\rho\cdot\frac{S_xS_y}{n}$$

---

[1] ลองกระจายกลับจะพบว่า $\frac{\bar{X}}{\bar{Y}}\left(1 + \frac{\bar{x}-\bar{X}}{\bar{X}}\right)\left(1 + \frac{\bar{y}-\bar{Y}}{\bar{Y}}\right)^{-1}$

$$= \frac{\bar{X}}{\bar{Y}}\cdot\frac{\bar{X}+\bar{x}-\bar{X}}{\bar{X}}\cdot\frac{\bar{Y}}{\bar{Y}+\bar{y}-\bar{Y}}$$

$$= \frac{\bar{X}}{\bar{Y}}\frac{\bar{x}}{\bar{X}}\frac{\bar{Y}}{\bar{y}} = \frac{\bar{x}}{\bar{y}}$$

โดยที่ ϱ คือสหสัมพันธ์ระหว่างตัวแปร x และตัวแปร y [1]

หรือ $\varrho = \varrho_{xy} = \varrho_{yx} = \dfrac{\Sigma(x-\bar{X})(y-\bar{Y})}{\sqrt{\Sigma(x-\bar{X})^2\Sigma(y-\bar{Y})^2}}$

$$S_x^2 = \frac{1}{N-1}\sum_{i}^{N}(x-\bar{X})^2,\ S_y^2 = \frac{1}{N-1}\sum_{i}^{N}(y-\bar{Y})^2$$

ดังนั้น $E(\hat{\bar{R}}) = R + \dfrac{N-n}{N}\cdot\dfrac{S_y^2}{n}\dfrac{R}{\bar{Y}^2} - \dfrac{N-n}{N}\varrho\dfrac{S_xS_y}{n}\cdot\dfrac{1}{\bar{X}\bar{Y}}\dfrac{\bar{X}}{\bar{Y}}$

: แทนที่ R ด้วย $\dfrac{\bar{X}}{\bar{Y}}$

$$= R + \frac{N-n}{N}\frac{S_y^2}{n}\frac{R}{\bar{Y}^2} - \frac{N-n}{N}\varrho\frac{S_xS_y}{n}\cdot\frac{1}{\bar{Y}^2}$$

ดังนั้น Bias คือ $E(\hat{\bar{R}}) - R = E(\hat{\bar{R}} - R)$

$$= \frac{N-n}{N}\cdot\frac{S_y^2}{n}\cdot\frac{R}{\bar{Y}^2} - \frac{N-n}{N}\varrho\frac{S_xS_y}{n}\cdot\frac{1}{\bar{Y}^2}$$

$$=> \quad E(\hat{\bar{R}} - R) = \frac{N-n}{Nn}\frac{1}{\bar{Y}^2}(RS_y^2 - \varrho S_xS_y)^{\ 2}$$

นั่นคือปริมาณของความเอียงเฉเท่ากับ $\dfrac{N-n}{Nn}\cdot\dfrac{1}{\bar{Y}^2}(RS_y^2 - \varrho S_xS_y)$

---

[1] การพิสูจน์ว่า $E(\bar{x}-\bar{X})(\bar{y}-\bar{Y}) = \dfrac{N-n}{nN}\varrho S_xS_y$ ได้แสดงไว้ท้ายตอนนี้แล้ว

[2] ถ้า $n\to\infty$ แล้ว $\dfrac{N-n}{Nn}\to 0$ เพราะตัวหารมีค่าสูง แสดงว่า ถ้าใช้กลุ่มตัวอย่างให้มีขนาดใหญ่แล้วปริมาณความเอียงเฉจะเข้าใกล้ 0 หรือ $E(\hat{\bar{R}} - R)\to 0$ หรือ หรือ $E(\hat{\bar{R}})\to R$

ซึ่ง $\hat{R}$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ R ก็ต่อเมื่อ

$$\frac{N-n}{Nn}\frac{1}{\bar{Y}^2}(RS_y^2 - \varrho S_x S_y) = 0$$

นั่นคือ ปริมาณของความเอียงเฉ่จะมีค่าเท่ากับ $\frac{N-n}{Nn}\frac{1}{\bar{Y}^2}(RS_y^2 - \varrho S_x S_y)$ ซึ่ง $\hat{R}$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ R ก็ต่อเมื่อ

$$\frac{N-n}{Nn}\frac{1}{\bar{Y}^2}(RS_y^2 - \varrho S_x S_y) = 0$$

หรือ $RS_y^2 = \varrho S_x S_y$

หรือ $RS_y = \varrho S_x$

การอภิปรายผลข้างต้นโดยอาศัยกราฟจะทำให้มองเห็นภาพที่ชัดเจนขึ้น ดังนี้

พิจารณาสมการถดถอย $X_i = \beta_0 + \beta_1 Y_i$ เมื่อ Y เป็นตัวแปรอิสระและ X เป็นตัวแปรตาม[1] ถ้าสมการ ผ่านจุดกำเนิดหรือจุดตัดบนแกน X เท่ากับ 0 แล้ว ความสัมพันธ์นี้จะลดรูปเป็น $X_i = \beta_1 Y_i$ และโดยอาศัยวิธี Least Square ทำให้ทราบว่า

$$\beta_1 = \frac{\sum_{i}^{N}(x_i-\bar{X})(y_i-\bar{Y})}{\sum_{i}^{N}(y_i-\bar{Y})^2}$$

---

[1] ก. โดยปกติเรานิยมเสนอสมการถดถอยในรูป $Y_i = \beta_0 + \beta_1 X_i$ ในที่นี้ใช้เป็น $X_i = \beta_0 + \beta_1 Y_i$ เพื่อให้สอดคล้องกับสูตร $V(\hat{R})$ ที่เสนอมาแล้ว ความจริงเราจะนิยามให้ใครเป็นตัวแปรอิสระหรือตัวแปรตามก็ได้

ข. อัตราส่วน R เป็นเรื่องของการเปรียบเทียบกันระหว่างตัวแปร ความสัมพันธ์ระหว่างตัวแปร เช่น ความถดถอยและสหสัมพันธ์ย่อมมีส่วนสำคัญในการอธิบายความหมายและช่วยทำให้เข้าใจการนำค่า R ไปใช้ประโยชน์ได้กระจ่างขึ้น

อนึ่ง จากสมการ $X_i = \beta_i Y_i$ ถ้ารวมตลอดถึง N ค่าแล้วหารด้วย N (ใส่ $\sum_i^N$ ทั้งสองด้านแล้วหารตลอดด้วย N) จะพบว่า

$$\frac{1}{N}\sum_i^N X_i = \frac{\beta_1}{N}\sum_i^N Y_i$$

$$=> \beta_1 = \bar{X}/\bar{Y}$$

แต่ $\frac{\bar{X}}{\bar{Y}} = R$ ดังนั้น $\beta_1 = R$ นั่นก็คือ $RS_y = \beta_1 S_y$

พิจารณา RS_y จะพบว่า

$$RS_y = \beta_1 S_y = \frac{\sum_i^N (x_i - \bar{X})(y_i - \bar{Y})}{\sum_i^N (y_i - \bar{Y})^2} \cdot S_y$$

$$= \frac{Cov(x,y)}{S_y^2} \cdot S_y = \frac{Cov(x,y)}{S_y}$$

$$= \frac{Cov(x,y)}{S_x S_y} \cdot S_x$$

$$= \rho S_x$$

แสดงว่า ถ้า $\beta_0 = 0$ หรือสมการถดถอยผ่านจุดกำเนิดแล้วจะมีผลให้ $RS_y = \rho S_x$ หรือนัยหนึ่ง ถ้าสมการความสัมพันธ์ระหว่างตัวแปร x และ y พุ่งผ่านจุดกำเนิดมาแล้ว $\hat{R}$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ R

ดังนั้น ในชั้นต้นเราควรนำค่าสังเกต (x,y) มาพล็อตกราฟดูแนวของจุดว่าเป็นแนวพุ่งผ่านจุดกำเนิดหรือไม่ ถ้าผ่านก็แสดงว่า ค่า $\hat{R} = \frac{\bar{x}}{\bar{y}}$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ R

อนึ่งเราสามารถใช้ประโยชน์จากข้อสนเทศเหล่านี้ในการคำนวณ $V(\hat{R})$ ได้อีกด้วย เพราะเราสามารถเสนอสูตรของ $V(\hat{R})$ ได้อีกรูปหนึ่งดังนี้

จาก $V(\hat{R}) = \frac{1}{\bar{Y}^2}\frac{N-n}{N}\frac{S_d^2}{n}$

$$= \frac{1}{\bar{Y}^2}\cdot\frac{N-n}{Nn}\cdot\frac{1}{N-1}\sum_i^N (x_i - Ry_i)^2$$

พิจารณาเทอม $\sum_i^N (x_i - Ry_i)^2$ จะพบว่า

$$\sum_i^N (x_i - Ry_i)^2 = \sum_i^N (x_i - \bar{X} + \bar{X} - Ry_i)^2$$

$$= \sum_i^N (x_i - \bar{X} + \bar{X}\cdot\frac{\bar{Y}}{\bar{Y}} - Ry_i)^2$$

$$= \sum_i^N (x_i - \bar{X} + R\bar{Y} - Ry_i)^2$$

$$= \sum_i^N \{(x_i - \bar{X}) - R(y_i - \bar{Y})\}^2$$

$$= \sum_i^N (x_i - \bar{X})^2 - 2R\sum_i^N (x_i - \bar{X})(y_i - \bar{Y}) + R^2\sum_i^N (y_i - \bar{Y})^2$$

$$= (N-1)S_x^2 - 2R(N-1)\varrho S_x S_y + (N-1)R^2S_y^2$$ [1]

---

[1] $\varrho = \frac{E(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{E(x_i - \bar{X})^2}\sqrt{E(y_i - \bar{Y})^2}} = \frac{\sum_i^N (x_i - \bar{X})(y_i - \bar{Y})}{(N-1)S_x S_y}$

ดังนั้น $\sum_i^N (x_i - \bar{X})(y_i - \bar{Y}) = (N-1)\varrho S_x S_y$

$$= (N-1)(S_x^2 + R^2S_y^2 - 2\rho RS_xS_y)$$

$$\Rightarrow \quad V(\hat{R}) = \frac{1}{\bar{Y}^2} \cdot \frac{N-n}{Nn} \cdot \frac{1}{N-1} \{(N-1)(S_x^2 + R^2S_y^2 - 2\rho RS_xS_y)$$

นั่นคือ $$V(\hat{R}) = \frac{1}{\bar{Y}^2} \cdot \frac{N-n}{Nn} \cdot (S_x^2 + R^2S_y^2 - 2\rho RS_xS_y)$$

พิจารณาสูตร $$V(\hat{R}) = \frac{1}{\bar{Y}^2} \frac{N-n}{Nn} (S_x^2 + R^2S_y^2 - 2R\rho S_xS_y)$$

จะพบว่าเนื่องจาก $-1 \leq \rho \leq 1$ ถ้า $\rho = 1$ แล้ว $S_x^2 + R^2S_y^2 - 2R\rho S_xS_y$ จะมีค่าต่ำที่สุด ดังนั้นเราจึงได้ข้อสรุปอันเป็นเงื่อนไขของ $\hat{R}$ ดังนี้คือ

1. ถ้าสมการ ความสัมพันธ์ระหว่างตัวแปร x และ y เป็นเส้นตรงที่พุ่งผ่านจุดกำเนิด แล้ว $\hat{R}$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ R

2. ถ้า x และ y มีค่าสหสัมพันธ์เป็น 1 (perfect positive correlation) ความแปรปรวนของการประมาณค่าจะมีค่าต่ำที่สุด

3. ปริมาณความเอียงเฉจะลดลงเมื่อใช้กลุ่มตัวอย่างขนาดใหญ่

4. ตัวประมาณค่าของ $V(\hat{R})$ ในกรณีของสูตรรูปนี้คือ

$$\hat{V}(\hat{R}) = \frac{1}{\bar{y}^2} \frac{N-n}{Nn} (s_x^2 + \hat{R}s_y^2 - 2\hat{R}rs_xs_y)$$

โดยที่ $$s_x^2 = \frac{1}{n-1} \sum_i^n (x_i - \bar{X})^2, \; s_y^2 = \frac{1}{n-1} \sum_i^n (y_i - \bar{Y})^2, \; \hat{R} = \bar{x}/\bar{y}$$

$r = s_{xy}/s_xs_y$ ค่าประมาณของสหสัมพันธ์ระหว่างตัวแปร x และ y

$$s_{xy} = \frac{1}{n-1} \sum_i^n (x_i - \bar{x})(y_i - \bar{y})$$

พิสูจน์ $E(\bar{x}-\bar{X})(\bar{y}-\bar{Y}) \stackrel{?}{=} \frac{N-n}{Nn}\varrho S_x S_y$

ให้ $u_i = x_i + y_i$ ดังนั้น $\bar{u} = \bar{x} + \bar{y}$ และ $\bar{U} = \bar{X} + \bar{Y}$

ดังนั้น $E(\bar{u}-\bar{U})^2 = E\{(\bar{x}+\bar{y})-(\bar{X}+\bar{Y})\}^2 = E\{(\bar{x}-\bar{X})+(\bar{y}-\bar{Y})\}^2$

$$= E(\bar{x}-\bar{X})^2 + E(\bar{y}-\bar{Y})^2 - 2E(\bar{x}-\bar{X})(\bar{y}-\bar{Y})$$

นั่นคือ $E(\bar{x}-\bar{X})(\bar{y}-\bar{Y}) = \frac{1}{2}\{(E(\bar{u}-\bar{U})^2 - E(\bar{x}-\bar{X})^2 - E(\bar{y}-\bar{Y})^2\}$

$$= \frac{1}{2}V(\bar{u}) - \frac{1}{2}V(\bar{x}) - \frac{1}{2}V(\bar{y})$$

$$= \frac{1}{2}\left(\frac{N-n}{N}\cdot\frac{S_u^2}{n} - \frac{N-n}{N}\cdot\frac{S_x^2}{n} - \frac{N-n}{N}\cdot\frac{S_y^2}{n}\right)$$

$$= \frac{N-n}{2Nn}(S_u^2 - S_x^2 - S_y^2)$$

โดยที่ $S_u^2 = \frac{1}{N-1}\sum_i^N (u_i-\bar{U})^2$

$$= \frac{1}{N-1}\sum_i^N \{(x_i+y_i)-(\bar{X}+\bar{Y})\}^2$$

$$= \frac{1}{N-1}\sum_i^N \{(x_i-\bar{X})+(y_i-\bar{Y})\}^2$$

$$= \frac{1}{N-1}\sum_i^N (x_i-\bar{X})^2 + \frac{1}{N-1}\sum_i^N (y_i-\bar{Y})^2 + \frac{2}{N-1}\sum_i^N (x_i-\bar{X})(y_i-\bar{Y})$$

$$= S_x^2 + S_y^2 + \frac{2}{N-1}\sum_i^N (x_i-\bar{X})(y_i-\bar{Y})$$

$$\text{ดังนั้น } E(x-\bar{X})(y-\bar{Y}) = \frac{N-n}{2Nn}\{(S_x^2+S_y^2+\frac{2}{N-1}\sum_i^N (x_i-\bar{X})(y_i-\bar{Y}))-S_x^2-S_y^2\}$$

$$= \frac{N-n}{Nn}\cdot\frac{1}{N-1}\sum_i^N (x_i-\bar{X})(y_i-\bar{Y})$$

$$\text{แต่ } \because \quad \varrho = \frac{\sum_i^N (x_i-\bar{X})(y_i-\bar{Y})}{(N-1)S_xS_y}$$

$$\sum_i^N (x_i-\bar{X})(y_i-\bar{Y}) = (N-1)\varrho S_xS_y$$

$$\text{ดังนั้น } E(\bar{x}-\bar{X})(\bar{y}-\bar{Y}) = \frac{N-n}{Nn}\cdot\frac{1}{N-1}\sum_i^N (x_i-\bar{X})(y_i-\bar{Y})$$

$$= \frac{N-n}{Nn}\varrho S_xS_y$$

2.4.3 การประมาณค่าสัดส่วนของกลุ่มประชากร (Estimation of Population Proportion or Population Percentage, p)

ในบางครั้งเรามีความสนใจที่จะทราบค่าประมาณของสัดส่วนหรือร้อยละของคุณลักษณะทางประชากรที่สนใจ หรือสนใจจะทราบค่าประมาณของยอดรวมของจำนวนหน่วยสำรวจที่มีลักษณะสอดคล้องกับคุณลักษณะทางประชากรที่เราสนใจ เช่น ร้อยละของผู้เป็นโรคผิวหนัง หรือยอดรวมของผู้ที่เป็นโรคผิวหนัง ร้อยละของนักเรียนสายตาสั้น หรือยอดรวมของนักเรียนที่สายตาสั้น ฯลฯ ในกรณีเช่นนี้เรามีวิธีประมาณค่าเรียกว่าการประมาณค่าสัดส่วนของกลุ่มประชากร

ในที่นี้เราจะเริ่มศึกษาเฉพาะกรณีที่กลุ่มประชากรถูกจำแนกเป็น 2 ฝ่าย (Class) คือมีคุณลักษณะตรงตามความสนใจของเราฝ่ายหนึ่ง เรียกว่าฝ่าย C (Class C) กับไม่มีคุณลักษณะตรงตามความสนใจของเราอีกฝ่ายหนึ่ง เรียกว่า C′ (Class C′) กล่าวเช่นนี้อาจทำให้มองภาพไม่ออก ขอยกตัวอย่างประกอบเพื่อให้เข้าใจง่ายดังนี้ สมมุติว่าเราสนใจที่จะ

กะประมาณสัดส่วนของนักศึกษาที่เราสนใจในกิจกรรมทางศาสนาและวัฒนธรรม ในที่นี้จะเห็นได้ว่าเราสามารถจำแนกนักศึกษาทั้งหมดออกได้เป็น 2 พวก หรือ 2 ฝ่ายคือ ฝ่ายหนึ่งเป็นฝ่ายที่มีความสนใจในกิจกรรมของศาสนาและวัฒนธรรม อีกฝ่ายหนึ่งไม่สนใจในกิจกรรมดังกล่าว ดั่งนี้เป็นต้น เมื่อเป็นเช่นนี้จึงเห็นได้ว่าถ้าดำเนินการสุ่มตัวอย่างหน่วยสำรวจขึ้นมา n หน่วย หน่วยสำรวจหนึ่ง ๆ จึงอาจตกอยู่ในฝ่าย C หรือ C' ก็ได้ หรือนัยหนึ่งหน่วยสำรวจหนึ่ง ๆ อาจมาจากฝ่าย C หรือ C' ก็ได้

ให้ $x = 1$ ถ้า $X \in C$ และ $X = 0$ ถ้า $X \in C'$ ดังนั้น เมื่อสุ่มตัวอย่างหน่วยสำรวจมา n หน่วยจากกลุ่มประชากรขนาด N เราสามารถนิยามสัดส่วนจากกลุ่มตัวอย่างและสัดส่วนของประชากรได้ดังนี้

นิยาม 2.1 เมื่อตัวแปรสุ่ม มีค่าได้เป็น 0 หรือ 1 อย่างใดอย่างหนึ่งเท่านั้นคือ $x = 0$ เมื่อ $x \in C'$ และ $x = 1$ เมื่อ $x \in C$ ดังนั้นสัดส่วนของประชากร P และสัดส่วนจากกลุ่มตัวอย่างจะมีลักษณะดังนี้คือ

$$P = \sum_{i}^{N} \frac{x_i}{N} \qquad \text{เมื่อ } x = 0,1$$

และ

$$p = \sum_{i}^{n} \frac{x_i}{n} \qquad \text{เมื่อ } x = 0,1$$

ข้อสังเกต

1. ขอให้สังเกตว่าทั้ง P และ p มีสูตรโครงสร้างเช่นเดียวกันกับ $\bar{X}$ และ $\bar{x}$ แตกต่างกันเฉพาะในกรณีของสัดส่วนค่าของ x มีได้เพียง 2 ค่าคือ 0 และ 1 เท่านั้นโดยที่ $x = 0$ เมื่อ $x \in C'$ และ $x = 1$ เมื่อ $x \in C$ ส่วนในกรณีของค่าเฉลี่ยค่าของ x จะมีได้ไม่จำกัดกล่าวคือ $-\infty < x < \infty$ ดังนั้นเราจึงสามารถสรุปได้ว่าสัดส่วนก็คือกรณีเฉพาะของค่าเฉลี่ยเมื่อ $x = 0,1$

2. ในทางปฏิบัติ หมายเลข 0 และ 1 ที่ใช้เป็นค่าของ x นั้นก็คือระหัสที่กำหนดให้แก่ x นั่นเอง เรานิยมกำหนดระหัสเช่นนี้ให้แก่ตัวแปรเชิงคุณภาพ เช่น เพศ สถานภาพสมรส

ความสัมพันธ์กับหัวหน้าครอบครัว ฯลฯ แต่จะกำหนดให้แก่ตัวแปรเชิงปริมาณด้วยก็ได้ ถ้าสามารถจแนกประชากรออกเป็น 2 กลุ่มที่มีความหมายได้

3. เหตุที่ตัวแปร x มีค่าได้ 2 ค่าคือ 0 และ 1 แสดงว่าฟังก์ชันการแจกแจงของตัวแปรสุ่มคือการแจกแจงแบบเบอร์นุลลิ (Bernoulli Distribution) และตัวสถิติ $y = \sum_{i}^{N} x_i$ จะมีการแจกแจงแบบทวินาม (Binomial Distribution).

4. $P = \sum_{i}^{N} \frac{x_i}{N}$ เมื่อ $x = 0$ หรือ 1 แสดงให้เห็นว่า ถ้า $x_i = 0$ แล้ว $x_i^2 = 0$ และถ้า $x_i = 1$ แล้ว $x_i^2 = 1$ ดังนั้นเราสามารถสรุปผลทางคณิตศาสตร์ที่น่าสนใจจากความเป็นจริงต่อไปนี้

(1) $\sum_{i}^{N} x_i^2 = \sum_{i}^{N} x_i$

(2) $P = \bar{X}$ = สัดส่วนของประชากรที่ตกอยู่ใน Class C

(3) $\sum_{i}^{N} x_i$ คือ ยอดรวมของประชากรที่ตกอยู่ใน Class C ถ้าให้ $\sum_{i}^{N} x_i = A$ จะสามารถเขียนสูตรใหม่ได้ดังนี้คือ

$$P = A/N$$

นั่นคือ $A = NP$ หรือ $\sum_{i}^{N} x_i = NP$ หรือ $\sum_{i}^{N} x_i = NP = N\bar{X}$

ในทำนองเดียวกัน จาก $p = \sum_{i}^{n} \frac{x_i}{n}$ ดังนั้น

(1) $\sum_{i}^{n} x_i^2 = \sum_{i}^{n} x_i$

(2) $p = \bar{x}$ = ค่าประมาณของสัดส่วนประชากรที่ตกอยู่ใน Class C

(3) ถ้าให้ $\sum_{i}^{n} x_i = a$ แสดงว่า $p = a/n$ นั่นคือ $a = np$ = ยอดรวมของหน่วยประชากรจาก Class C ที่ตกอยู่ในกลุ่มตัวอย่างขนาด n

ทฤษฎี 2.5 ถ้าสุ่มตัวอย่างหน่วยสำรวจมา n ชุดจากกลุ่มประชากรขนาด N โดยใช้แผนสำรวจแบบ SRS แล้ว

1. $\hat{P} = p = \sum_{i}^{n} \frac{x_i}{n}$ = จะใช้เป็นตัวประมาณค่าของ $P = \sum_{i}^{N} \frac{x_i}{N}$

2. ความแปรปรวนของค่าประมาณ $\hat{P}$ คือ

$$V(\hat{P}) = \frac{N-n}{N-1} \cdot \frac{PQ}{n}$$

3. ตัวประมาณค่าที่ปราศจากความอคติของ $V(\hat{P})$ คือ

$$\hat{V}(\hat{P}) = \frac{N-n}{N} \cdot \frac{pq}{n-1}$$

พิสูจน์

1. $E(\hat{P}) \stackrel{?}{=} P$

$$E(\hat{P}) = E\left(\frac{1}{n} \sum_{i}^{n} x_i\right)$$

$$= \frac{1}{n} \sum_{i}^{n} E(x_i)$$

$$= \frac{1}{n} \sum_{i}^{n} \left(\sum_{i}^{N} x_i \cdot \frac{1}{N} \right)$$

$$= \frac{1}{n} n \cdot \left(\frac{1}{N} \sum_{i}^{N} x_i \right) = P$$

แสดงว่า $\hat{P} = p = \sum_{i}^{n} \frac{x_i}{n}$ เป็นตัวประมาณค่าที่ปราศจากอคติของ P และเราสามารถใช้ค่า $\hat{p}$ เป็นค่าประมาณของ P ได้

2. เนื่องจาก $\hat{P} = p$ เป็นกรณีเฉพาะของ $\bar{x}$ โดย $x_i = 0, 1$ ดังนั้นโครงสร้างของสูตรความแปรปรวนของค่าประมาณ $V(\hat{P})$ จึงเป็นโครงสร้างเดียวกัน

ดังนั้น

$$V(\hat{P}) = \frac{N-n}{N}\frac{S^2}{n}$$

พิจารณา S^2 สำหรับกรณีของสัดส่วนจะพบว่า

$$S^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i-\bar{X})^2$$

$$= \frac{1}{N-1}\sum_{i}^{N}(x_i^2-2x_i\bar{X}+\bar{X}^2)$$

$$= \frac{1}{N-1}\left(\sum_{i}^{N}x_i^2-N\bar{X}^2\right)$$

แต่ $\sum_{i}^{N}x_i^2 = \sum_{i}^{N}x_i = NP$ และ $\bar{X}^2 = P^2$ (ดูข้อสังเกตท้ายนิยาม 2.1)

$$=> S^2 = \frac{1}{N-1}(NP-NP^2)$$

$$= \frac{1}{N-1}\cdot NP(1-P)$$

$$= \frac{NPQ}{N-1} \quad ; Q = 1-P$$

ดังนั้น $V(\hat{P}) = \frac{N-n}{N}\frac{S^2}{n} = \frac{N-n}{N}\frac{1}{n}\cdot\frac{NPQ}{N-1}$

$$=> V(\hat{P}) = \frac{N-n}{N-1}\frac{PQ}{n}$$

แต่เนื่องจาก P และ Q เป็นพารามิเตอร์ซึ่งเป็นตัวที่ไม่ทราบค่า ดังนั้นในทางปฏิบัติจึงไม่อาจใช้สูตรดังกล่าวนี้ในการคำนวณหาความแปรปรวนของค่าประมาณได้แต่ให้ใช้สูตรต่อไปนี้แทนคือ

$$\hat{V}(\hat{P}) = \frac{N-n}{N}\frac{pq}{n-1}$$

3. จาก $V(\hat{P}) = \frac{N-n}{N} \cdot \frac{S^2}{n}$

แต่เนื่องจาก $E(s^2) = S^2$ หรือ s^2 เป็นตัวประมาณค่าที่ปราศจากอคติของ S^2 ดังนั้นในทางปฏิบัติจึงใส่ s^2 ลงไปแทนที่ของ S^2 ทำให้ได้ค่าประมาณของ $V(\hat{P})$ เป็น

$$\hat{V}(\hat{P}) = \frac{N-n}{N}\frac{s^2}{n}$$

แต่ $s^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i-\bar{x})^2$

$$= \frac{1}{n-1}\left(\sum_{i}^{n} x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i}^{n} x_i - np^2\right)$$

$$= \frac{1}{n-1}(np - np^2)$$

$$= \frac{np(1-p)}{n-1}$$

$$= \frac{npq}{n-1}$$

ดังนั้น $\hat{V}(\hat{P}) = \frac{N-n}{N}\frac{s^2}{n} = \frac{N-n}{N} \cdot \frac{1}{n} \cdot \frac{npq}{n-1}$

$=> \hat{V}(\hat{P}) = \frac{N-n}{N}\frac{pq}{n-1}$

บทแทรก 2.3 เมื่อสุ่มตัวอย่างขนาด n มาจากกลุ่มประชากร N โดยแผนสำรวจแบบ SRS เพื่อกะประมาณสัดส่วน

1. $\hat{T} = N\hat{P} = Np$ เป็นค่าประมาณที่ปราศจากอคติของยอดรวมของประชากรที่อยู่ใน Class C

2. $V(\hat{T}) = N^2V(\hat{P}) = N^2\frac{N-n}{N-1}\frac{PQ}{n}$

3. $\hat{V}(\hat{T}) = N^2\hat{V}(\hat{P}) = N^2\frac{N-n}{N-1}\frac{pq}{n-1}$

4. ช่วงเชื่อมั่น $(1-\alpha)\,100\%$ ที่คาดว่าค่าจริงของ P จะปรากฏอยู่คือ

$$\hat{P} - Z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n-1}} < P < \hat{P} + Z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n-1}}$$

5. ช่วงเชื่อมั่น $(1-\alpha)\,100\%$ ที่คาดว่าค่าจริงของ T จะปรากฏอยู่คือ

$$Np + Z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n-1}} < T < Np + Z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n-1}}$$

พิสูจน์ ในที่นี้จะแสดงการพิสูจน์เฉพาะข้อ 1 เท่านั้นข้ออื่น ๆ จะขอเว้นไว้เป็นแบบฝึกหัด ซึ่งนักศึกษาสามารถพิสูจน์ได้เองโดยง่าย

1. $E(\hat{T}) = E(Np)$

$= NE(p)$

$= NP \quad (\because E(p) = E(\hat{P}) = P)$

$= N \cdot \frac{1}{N}\sum_{i}^{N} x_i$

$= \sum_{i}^{N} x_i = T$

แสดงว่า $\hat{T} = Np$ เป็นตัวประมาณค่าที่ปราศจากอคติของยอดรวมประชากร

หมายเหตุ

1. ถ้า $n < 30$ ให้ใช้ $t_{n-1,1-\alpha/2}$ และ $t_{n-1,\alpha/2}$ แทน $Z_{1-\alpha/2}$ และ $Z_{\alpha/2}$ ตามลำดับ

2. $Z_{\alpha/2} = -Z_{1-\alpha/2}$ เช่น กำหนดให้ $\alpha = 5\%$ ดังนั้น $Z_{1-\alpha/2} = Z_{.975} = 1.96$ และ $Z_{\alpha/2} = Z_{.025} = -1.96$ จะเห็นได้ว่า $Z_{.025} = -1.96 = Z_{.975}$

ดังนั้นทางซ้ายของอสมการในข้อ 4 และ 5 เราจะใช้ $\hat{P} - Z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n-1}}$

และ $Np - Z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n-1}}$ ก็ได้ อนึ่ง $\hat{P}$ และ p มีความหมายเดียวกันจะเขียนโดยใช้ตัวใดก็ได้

3. ดังที่ได้กล่าวมาแล้วในตอนต้นว่า สำหรับกรณีของการประมาณค่าสัดส่วนนั้น เราถือว่าตัวแปรสุ่ม x มีค่าเพียง 2 ค่าคือ $x = 1$ เมื่อ $x \in C$ และ $x = 0$ เมื่อ $x \in C'$ ดังนั้นในทางทฤษฎีตัวแปรสุ่ม x จะมีการกระจายแบบเบอร์นุลลิคือ

$$f(x) = p^x(1-p)^{1-x} \quad ; \quad x = 0, 1$$

และเมื่อสุ่มตัวอย่างมา n หน่วยตัวแปรสุ่ม $Y = x_1 + x_2 + \ldots + x_n$ จะมีการกระจายแบบทวินามคือ

$$f(y) = \binom{n}{y} p^y q^{n-y} \quad ; \quad y = 0, 1, \ldots, n$$

ปัญหาที่พบก็คือ ตัวสถิติ $p = \sum_i^n \frac{x_i}{n}$ หรือ $\frac{y}{n}$ มีการแจกแจงแบบใดเพราะถ้าไม่ทราบ Sampling Distribution ของ p เราจะไม่สามารถประมาณค่าของ P ด้วยช่วงเชื่อมั่นหรือทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ P ได้[1]

---

[1] อ่านมนตรี พิริยะกุล **ทฤษฎีสถิติ 2** (โรงพิมพ์มหาวิทยาลัยรามคำแหง, กทม. 2520) บทที่ 2

จากการศึกษาเรื่องทฤษฎีการโน้มเข้าสู่เกณฑ์กลาง (Central Limit Theorem) เราสามารถพิสูจน์ได้ว่า ถ้า $n \to \infty$ แล้วตัวแปรสุ่ม $p = \frac{\sum_{i}^{n} x_i}{n}$ จะมีการแจกแจงแบบปกติโดยประมาณ มีค่าเฉลี่ย $\mu = P$ และความแปรปรวน $\sigma^2 = \frac{N-n}{N-1}\frac{PQ}{n}$

หรือ $p \sim N(P, \frac{N-n}{N-1}\frac{PQ}{n})$

ด้วยความจริงประการดังกล่าวจึงทำให้เราสามารถกะประมาณค่าของพารามิเตอร์ P ด้วยช่วงเชื่อมั่น $(1-\alpha)\ 100\%$ ได้ดังบทแทรก 2.3

ปัญหาอีกประการหนึ่งก็คือ ตามปกติถ้า $p \simeq .5$ แล้ว Sampling Distribution ของ p จึงจะเป็นแบบปกติได้โดยประมาณ แต่ถ้า $p \to 0$ หรือ $p \to 1$ แล้ว Sampling Distribution ของ p จะเบ้ไม่มีลักษณะเป็นแบบปกติได้ ปัญหาจึงขึ้นอยู่กับขนาดตัวอย่าง n ว่าควรจะเป็นเท่าไร จึงจะแก้ปัญหาโค้งเบ้ได้

เกี่ยวกับปัญหานี้เรามีทางแก้ปัญหาของการคำนวณกำหนดขนาดตัวอย่างได้โดยใช้ขนาดตัวอย่างขั้นต่ำตามตารางซึ่งเสนอโดย W.G. Cochran ดังนี้

ค่าของ p	ค่าขนาดตัวอย่าง n ขั้นต่ำ
0.5	30
0.4 หรือ 0.6	50
0.3 หรือ 0.7	80
0.2 หรือ 0.8	200
0.1 หรือ 0.9	600
0.05 หรือ 0.95	1400