บทที่ 2
สถิติศาสตร์และกระบวนการทางสถิติ

2.1 ความนำ

สถิติศาสตร์ถูกนำมาใช้ในการหาข้อสรุปเกี่ยวกับสิ่งที่สนใจ โดยการใช้ข้อมูลเพียงบางส่วนของสิ่งที่สนใจนั้นมาทำการอนุมานทางสถิติ เพื่อให้ได้ผลสรุปและสามารถอธิบายเกี่ยวกับลักษณะของสิ่งที่สนใจ สถิติศาสตร์สามารถแบ่งได้เป็น 2 ส่วน คือ สถิติเชิงพรรณนา (Descriptive Statistics) เป็นการสรุปเกี่ยวกับข้อมูลชุดหนึ่งที่สนใจ โดยไม่มีการอ้างอิงถึงประชากร และสถิติเชิงอนุมาน (Inferential Statistics) เป็นการนำข้อมูลตัวอย่างมาทำการสรุปเกี่ยวกับประชากรที่สนใจทั้งประชากร ทฤษฎีที่เกี่ยวกับการอธิบายลักษณะประชากร คือ ทฤษฎีการอนุมานทางสถิติ และ ทฤษฎีการสำรวจตัวอย่าง ใช้สำหรับการอธิบายกลุ่มประชากรที่สนใจ

2.2 ทฤษฎีการอนุมานทางสถิติ (Theory of Statistical Inference)

การอนุมาน (Inference) เป็นการศึกษาเกี่ยวกับการนำข้อมูลตัวอย่างไปอธิบายถึงข้อมูลทั้งหมดของทั้งประชากร การอนุมานทางสถิติ (Statistical Inference) เป็นการสร้างตัวแบบทางสถิติจากข้อมูลตัวอย่าง จากนั้นจึงนำตัวแบบที่ได้ไปทำการอนุมานสรุปเกี่ยวกับประชากรที่กำลังศึกษา โดยคำนึงถึงคุณภาพของการเป็นตัวแทนที่ดีของประชากรจากตัวอย่างที่สุ่มได้ การอนุมานทางสถิตินั้นแยกออกได้เป็น 3 ลักษณะ คือ การประมาณค่า (Estimation) การทดสอบสมมติฐาน (Hypothesis Testing) และการตัดสินใจ (Decision)

การประมาณค่า คือ การใช้ข้อมูลจากตัวอย่างในการประมาณค่าลักษณะหรือพารามิเตอร์ต่าง ๆ ของประชากร การประมาณค่าสามารถกระทำได้ด้วยการประมาณค่าแบบจุด (Point Estimation) และการประมาณค่าแบบช่วง (Interval Estimation) การทดสอบสมมติฐาน คือ การใช้ข้อมูลจากตัวอย่างและความรู้เกี่ยวกับตัวแบบความน่าจะเป็น มาทดสอบความเชื่อเกี่ยวกับค่าพารามิเตอร์หรือประชากรที่สนใจ และการตัดสินใจ คือ การใช้ข้อมูลจากตัวอย่างมาช่วยในการตัดสินใจเลือกการกระทำ (Action) ที่เหมาะสมเพื่อให้เกิดความสูญเสียที่เกิดจากการตัดสินใจเลือกการกระทำนั้นน้อยที่สุด

การสร้างตัวประมาณ

ตัวประมาณ (Estimator) คือ ฟังก์ชันที่ใช้ประมาณค่าพารามิเตอร์ โดยฟังก์ชันที่ใช้จะเป็นฟังก์ชันของตัวแปรในตัวอย่างสุ่ม ดังนั้นตัวประมาณ ก็คือ ตัวสถิติที่นำมาประมาณค่าพารามิเตอร์ ซึ่งสามารถหาตัวประมาณค่าได้หลายตัวประมาณสำหรับการประมาณค่าพารามิเตอร์หนึ่ง ๆ

การพิจารณาเลือกตัวสถิติที่นำมาเป็นตัวประมาณค่าพารามิเตอร์นั้น ต้องพยายามหาตัวสถิติที่สามารถให้ค่าประมาณใกล้เคียงกับค่าที่แท้จริงของพารามิเตอร์มากที่สุด โดยการพิจารณาคุณสมบัติที่ดีของตัวสถิติ ดังนี้

- ความไม่เอนเอียง (Unbiasedness)
- ความคงเส้นคงวา (Consistency)
- ความมีประสิทธิภาพ (Efficiency)
- ความเพียงพอ (Sufficiency)

วิธีการสร้างตัวประมาณมีหลายวิธี แต่วิธีหนึ่งที่ใช้กันมากคือ วิธีภาวะน่าจะเป็นสูงสุด (Method of Maximum Likelihood) เนื่องจากวิธีการนี้จะให้ตัวประมาณค่าที่มีคุณภาพที่ดีหลายประการ

การสร้างตัวประมาณด้วยวิธีภาวะน่าจะเป็น

แนวคิดของวิธีการนี้มีอยู่ว่า การประมาณค่าพารามิเตอร์ทำโดยอาศัยผลที่วัดได้จากตัวอย่างสุ่มที่เลือกมาจากการแจกแจงที่ทราบรูปแบบของฟังก์ชันความหนาแน่น แต่ไม่ทราบค่าพารามิเตอร์ จึงน่าจะใช้โอกาสที่เราจะเลือกตัวอย่างและวัดค่าได้ $(X_1 = x_1,\ldots,X_n = x_n)$ มาพิจารณาหาค่าประมาณของพารามิเตอร์ θ โอกาสที่จะวัดค่าตัวอย่างสุ่มได้ $X_1 = x_1,\ldots,X_n = x_n$ อาจแสดงได้ด้วยฟังก์ชันความหนาแน่นร่วมของ $X_1 = x_1,\ldots,X_n = x_n$ แต่ฟังก์ชันความหนาแน่นร่วมนี้ขึ้นอยู่กับค่าพารามิเตอร์ θ ดังนั้นค่าประมาณของพารามิเตอร์ θ ที่น่าจะได้รับการพิจารณา คือ ค่าของพารามิเตอร์ θ ที่ทำให้ฟังก์ชันความหนาแน่นร่วมนี้มีค่าสูงสุด

ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของตัวอย่างสุ่ม คือ ฟังก์ชันความหนาแน่นร่วมของ $X_1,\ldots,X_n$ โดยถือว่าเป็นฟังก์ชันของพารามิเตอร์ θ เรามักแทนฟังก์ชันภาวะน่าจะเป็นด้วย $L(x_1,\ldots,x_n;\theta)$ หรือ $L(\theta)$ นั่นคือ

$$L(x_1,\ldots,x_n;\theta) = f(x_1;\theta)\ldots f(x_n;\theta)$$

หลักในการสร้างตัวประมาณภาวะน่าจะเป็นสูงสุด มีดังนี้

ให้ $X_1,\ldots,X_n$ เป็นตัวอย่างสุ่มจากการแจกแจงที่มีฟังก์ชันความหนาแน่น $f(x;\theta)$ ซึ่งฟังก์ชันการแจกแจงความหนาแน่นร่วมของ $X_1,\ldots,X_n$ คือ

$$f(x_1,\ldots,x_n;\theta) = f(x_1;\theta)\ldots f(x_n;\theta)$$

โดยที่ฟังก์ชันความหนาแน่นร่วมของ $X_1,...,X_n$ อาจถือว่าเป็นฟังก์ชันของพารามิเตอร์ θ ซึ่งถ้าเป็นฟังก์ชันของพารามิเตอร์ θ จะเรียกฟังก์ชันนี้ว่า ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function)

การประมาณค่าพารามิเตอร์ θ ด้วยวิธีการประมาณแบบภาวะน่าจะเป็นสูงสุด คือการหาค่าของของพารามิเตอร์ θ ที่ทำให้ $L(x_1,...,x_n;\theta)$ มีค่ามากที่สุด

2.3 ทฤษฎีการสำรวจตัวอย่าง (Theory of Sample Survey)

ทฤษฎีการสำรวจตัวอย่างเป็นทฤษฎีทางสถิติที่พัฒนาขึ้นมาเพื่อใช้ในการอธิบายประชากรอันตะ แนวคิดพื้นฐานที่สำคัญ คือ ประชากรที่ศึกษาจะต้องนับจำนวนได้ ทฤษฎีนี้ว่าด้วยการเลือกตัวอย่างจากประชากรและการหาค่าประมาณจากตัวอย่าง เพื่อประมาณค่าคุณลักษณะของประชากรอย่างมีคุณภาพที่สุด ภายใต้ข้อจำกัดทางด้านทรัพยากร

การสร้างตัวประมาณ

หลักพื้นฐานในการสร้างตัวประมาณ คือ การปรับค่าในระดับตัวอย่างไปสู่ระดับประชากร โดยอาศัยความน่าจะเป็นที่ชุดตัวอย่างถูกเลือกมาจากประชากรนั้น ในทฤษฎีการสำรวจตัวอย่างไม่มีวิธีการสร้างตัวประมาณที่ชัดเจน

การวัดคุณภาพของตัวประมาณ

คุณภาพของค่าประมาณลักษณะประชากรแสดงถึงความน่าเชื่อถือของระเบียบวิธีทางสถิติที่ใช้ในการเก็บข้อมูล การวัดคุณภาพของข้อมูลพิจารณาได้ 2 เกณฑ์ คือ เกณฑ์แรกเป็นการวัดคุณภาพในลักษณะที่พิจารณาว่าค่าต่าง ๆ ที่เป็นไปได้ของตัวประมาณแตกต่างจากค่าประชากรหรือค่าจริงเพียงไร เรียกเกณฑ์นี้ว่า ความถูกต้อง (Accuracy) ของตัวประมาณ ตัวประมาณใดมีค่าที่เป็นไปได้ต่าง ๆ ใกล้เคียงค่าจริงมากกว่า ก็ย่อมจะถูกต้องกว่าตัวประมาณอื่นที่มีค่าที่เป็นไปได้แตกต่างห่างออกไป สำหรับค่าที่ใช้วัดความถูกต้องนั้น มักใช้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error ; MSE) เกณฑ์นี้ต้องพิจารณาความแตกต่างจากค่าประชากรหรือค่าจริงที่ไม่ทราบค่า ดังนั้นในทางปฏิบัติไม่สามารถกระทำได้ เกณฑ์ที่สองเป็นการวัดคุณภาพในลักษณะที่พิจารณาว่าค่าต่าง ๆ ที่เป็นไปได้ของตัวประมาณแตกต่างจากค่าคาดหวังของตัวประมาณนั้นเพียงไร เรียกเกณฑ์นี้ว่า ความแม่นยำ (Precision) ของตัวประมาณ ตัวประมาณใดมีค่าที่เป็นไปได้ต่าง ๆ ใกล้เคียงกับค่าคาดหวังของตัวประมาณมากกว่า ก็ย่อมจะแม่นยำกว่าตัวประมาณอื่นที่มีค่าที่เป็นไปได้แตกต่างห่างออกไป สำหรับค่าที่ใช้วัดความแม่นยำนั้น ใช้ความแปรปรวน (Variance) ของตัว

ประมาณ เกณฑ์การวัดคุณภาพทั้งสองเกณฑ์มีความสัมพันธ์กันคือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองมีค่าเท่ากับค่าความแปรปรวนของตัวประมาณ ถ้าตัวประมาณนั้นเป็นตัวประมาณที่ไม่เอนเอียง

2.4 วรรณกรรมที่เกี่ยวข้อง

David Nelson และ Glen Meeden ; ศึกษาเกี่ยวกับการประมาณค่าเฉลี่ยของประชากรเมื่อทราบเบื้องต้นว่า มัธยฐานของประชากรขึ้นอยู่กับช่วงใดช่วงหนึ่ง และศึกษาเกี่ยวกับการประมาณค่าเฉลี่ยของประชากรเมื่อทราบช่วงของมัธยฐานของตัวแปรช่วย (Auxiliary Variable) โดยแสดงการหา Polya Posterior ที่ใช้ในการแก้ปัญหาเมื่อทราบข้อมูลเบื้องต้นเพียงเล็กน้อย โดยที่ Polya Posterior เป็นการหาการแจกแจงร่วมของค่าที่ไม่ได้ถูกสังเกตของประชากรขนาด N-n

Edward L. Korn and Barry I. Graubard ; ได้ศึกษาเกี่ยวกับตัวประมาณความแปรปรวนของค่าประมาณของพารามิเตอร์ในอภิประชากรของตัวแบบสโตแคสติก (Stochastic Model) ภายใต้การจำลองค่าของประชากรอันตะ โดยพิจารณาความเหมาะสมในการใช้งานของค่าปรับประชากรอันตะ (finite population correction factors ; fpc factors) ซึ่งจะเห็นว่าการคำนวณความแปรปรวนของข้อมูลตัวอย่าง (Sampled data) ที่รวมค่าปรับประชากรอันตะไม่เหมาะสมสำหรับการนำไปใช้ประโยชน์ โดยพิจารณาจากคุณสมบัติความแกร่งของตัวประมาณ ในทางปฏิบัติการสุ่มตัวอย่างแบบง่ายจะไม่พิจารณาค่าปรับประชากรอันตะในการประมาณค่าความแปรปรวน ผลที่ได้นั้นจะเหมาะสมกับการอนุมานในอภิประชากรภายใต้ตัวแบบอภิประชากรแบบง่าย โดยการใช้ข้อมูลของการสำรวจสภาวะสุขภาพอนามัย (National Health Interview Survey) ปี 1987 เป็นกรณีศึกษา

สำหรับประชากรอันตะ ให้ $(y_1,\eta_1),\ldots,(y_K,\eta_K)$ เป็นค่าสังเกต โดยที่ y_i เป็นค่าจริงของตัวแปรสุ่มที่มีค่าเฉลี่ย m_i และความแปรปรวน t_i^2 ตัวแปรสุ่ม $Y_1,\ldots,Y_K$ เป็นอิสระซึ่งกันและกัน และ η_i เป็นตัวแปรแสดงชั้นภูมิที่นำไปใช้ในแผนการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ ดังนั้นเวคเตอร์พารามิเตอร์ 3 มิติ (dimension) ของ (m_i,t_i^2,η_i) เป็นอิสระซึ่งกันและกัน และมีการแจกแจงเหมือนกัน ซึ่งการแจกแจงจะมีลักษณะเช่นเดียวกันกับการแจกแจงของเวคเตอร์สุ่ม (μ,σ^2,η) คือการแจกแจงแบบเอฟ (F Distribution)

โดยทั่วไปนั้น ค่าเฉลี่ยของประชากรในประชากรอันตะจะเป็นค่าเฉลี่ยของค่าสังเกตในประชากร ส่วนค่าเฉลี่ยของประชากรในอภิประชากรจะอยู่ในรูป $\mu_{SP}=E_F(\mu)$

สัญลักษณ์ต่าง ๆ ในงานวิจัยนี้คือ

y_i หมายถึง ค่าของตัวอย่างที่ i โดยที่ $i=1,\ldots,k$

f หมายถึง ค่าสัดส่วนการสุ่ม (Sampling fraction ; $f=\frac{k}{K}$)

$1-f$ หมายถึง ค่าปรับประชากรอันตะ ((finite population correction factors ; fpc factors)

k หมายถึง ขนาดตัวอย่าง

K หมายถึง ขนาดประชากรทั้งหมดที่สนใจศึกษา

K_h หมายถึง ขนาดประชากรในชั้นภูมิ h

k_h หมายถึง ขนาดของตัวอย่างในชั้นภูมิ h โดยที่ $k_h = c_h(K_h)$ ซึ่งฟังก์ชัน c_h ขึ้นอยู่กับชั้นภูมิ h

$\bar{y}$ หมายถึง ค่าเฉลี่ยของตัวอย่าง

$\bar{y}_h$ หมายถึง ค่าเฉลี่ยของตัวอย่างในชั้นภูมิ h

σ^2 หมายถึง ความแปรปรวนตัวอย่าง

σ_h^2 หมายถึง ความแปรปรวนตัวอย่างในชั้นภูมิ h

ดังนั้นตัวประมาณความแปรปรวนของค่าเฉลี่ย $\bar{y}$ ในอภิประชากร คือ

กรณีการสุ่มตัวอย่างแบบง่าย ;

$$Var(\bar{y}) = E\left\{\hat{Var}_{wr}(\bar{y})\right\} = [E_F(\sigma^2) + Var_F(\mu)]/k$$

กรณีการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ ;

$$\hat{Var}_{SP}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h(K_h-1)}{K(K-1)} \frac{1}{k_h} \sigma_h^2 + \frac{1}{K-1}\left[\sum_{h=1}^{L} \frac{K_h}{K}\bar{y}_h^2 - \bar{y}^2\right]$$

V.R. Padmawar ; ศึกษาเปรียบเทียบตัวประมาณค่าเฉลี่ยของประชากรเมื่อทราบฟังก์ชันของตัวแบบถดถอยของบ่ระชากร ซึ่งได้แก่ ตัวประมาณโดยใช้อัตราส่วน (Ratio estimator ; t_R) , ตัวประมาณฮอร์วิทซ-ทอมป์สัน (Horvitz-Thomson estimator ; t_{HT}) , ตัวประมาณที่กำหนดค่าคงที่ g (estimator given by $\frac{\mu}{\sum x_i^{2-g}} \sum x_i^{1-g} y(x_i)$, $g \in [0,2]$; t_g) และตัวประมาณของราว – ฮาร์ทลีย์ - คอคครัน (Rao-Hartley-cochran estimator ; t_{RHC}) ซึ่งจากการศึกษาพบว่า ตัวประมาณ t_g เป็นตัวประมาณที่ดีที่สุดภายใต้เงื่อนไขความแปรปรวนต่ำสุดเนื่องจากตัวประมาณ t_g เป็นตัวประมาณที่ไม่เอนเอียง

พิจารณาประชากรอนันต์ (Infinite Population) $(y(x),x);x\geq 0$ ที่ทราบฟังก์ชันการแจกแจงความน่าจะเป็นสะสมร่วม (Joint distribution) ของ $y(x)$ (ξ) สำหรับฟังก์ชันการแจกแจงของ X คือ

$$F(x)=\int_0^x f(u)du\,;x\geq 0$$

โดยที่ Y เป็นตัวแปรที่สนใจศึกษา และ X เป็นตัวแปรช่วย (auxiliary variable)

ให้ $p(x)$ เป็นฟังก์ชันความน่าจะเป็นของแผนแบบ (design function) จะเรียกฟังก์ชัน t ของค่าสังเกต $(y(x),x)$ ว่าตัวประมาณค่าเฉลี่ยของประชากร m_Y โดยที่

$$m_Y=E_f(y)=\int_0^\infty y(x)f(x)dx$$

ซึ่งตัวแบบอภิประชากรที่เป็นตัวแบบถดถอย คือ

$$Y(x)=\beta x+Z(x)\quad ;x\geq 0$$

โดยที่ $E_\xi(Z(x))=0\quad ,E_\xi(Z^2(x))=\sigma^2x^g$

และ $E_\xi(Z(x_i)Z(x_j))=0\quad ;i\neq j$

เมื่อ $\sigma^2>0$, β ไม่ทราบค่า และ $g\in[0,2]$ ซึ่งอาจทราบหรือไม่ทราบก็ได้

ในการศึกษาของ V.R. Padmawar ได้ทำการเปรียบเทียบแผนแบบ (Strategies) $(srs,\bar{y})$, (srs,t_R) , (p_M,t_R) , (ppx,t_{HT}) , (p_g,t_g) และ (p_{RHC},t_{RHC}) โดยที่แผนแบบ (srs,t_R) เป็นแผนแบบที่มีความเอนเอียง ในขณะที่แผนแบบ $(srs,\bar{y})$, (ppx,t_{HT}) , (p_g,t_g) และ (p_{RHC},t_{RHC}) เป็นแผนแบบที่ไม่มีความเอนเอียง

เมื่อ srs คือแผนการสุ่มตัวอย่างแบบง่ายที่ค่า $p(x)\equiv\frac{1}{n}$

ppx^a คือแผนการสุ่มตัวอย่างที่ค่า $p(x)\propto\prod_{i=1}^{n}x_i^a$

p_M คือแผนการสุ่มตัวอย่างของมิดซูโน-เซน (Midzuno-Sen) ที่ค่า $p(x)=\frac{1}{n\mu}\sum x_i$

โดยที่ $\mu=\int_0^\infty xf(x)dx$

p_g คือแผนการสุ่มตัวอย่างที่ค่า $p(x)=k\prod_{i=1}^{n}x_i^{g-1}\sum x_i^{2-g}$

โดยที่ $k = \frac{1}{n\mu}[\frac{\Gamma(\alpha)}{\Gamma(\alpha+g-1)}]^{n-1} \quad ;(\mu=\alpha)$

p_{RHC} คือแผนการสุ่มตัวอย่างของราว – ฮาร์ทลีย์ - คอคครัน (Rao-Hartley-cochran)

$\bar{y}$ คือค่าเฉลี่ยของตัวอย่าง (Sample mean) $\bar{y} = \frac{1}{n}\sum y(x_i)$

t_R คือตัวประมาณอัตราส่วน (Ratio estimator) $t_R = \mu\frac{\sum y(x_i)}{\sum x_i}$

t_{HT} คือตัวประมาณฮอร์วิทซ-ทอมป์สัน (Horvitz-Thomson estimator)

$$t_{HT} = \sum \frac{y(x_i)f(x_i)}{\pi(x_i)}$$

t_g คือตัวประมาณที่กำหนดค่าคงที่ g

$$t_g = \frac{\mu}{\sum x_i^{2-g}}\sum x_i^{1-g} y(x_i)\,, g \in [0,2]$$

t_{RHC} คือตัวประมาณราว – ฮาร์ทลีย์ - คอคครัน (Rao-Hartley-cochran estimator)

ผลการศึกษาที่ได้คือ ตัวประมาณ t_g เป็นตัวประมาณที่ดีที่สุดสำหรับการประมาณค่าเฉลี่ยของประชากร ภายใต้เงื่อนไขของความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) ต่ำสุด และสำหรับค่าคงที่ g เมื่อกำหนดค่าคงที่ $g=1$ จะได้ว่าแผนแบบ (p_g, t_g) จะเหมือนกับแผนแบบ (p_M, t_R) และเมื่อกำหนดค่าคงที่ $g=2$ แผนแบบ (p_g, t_g) จะเหมือนกับแผนแบบ (ppx, t_{HT})