บทที่ ๒

ลักษณะของปัญหาทางสถิติ (Nature of Statistical Problems)

๒.๑ คำนำ

ปัญหาทางสถิติที่เราประสบอยู่ขณะนี้อาจจำแนกได้เป็นสองปัญหาใหญ่ ๆ ด้วยกัน อันได้แก่ปัญหาทางด้านการประมาณค่า (Estimation problems) เป็นการประมาณค่าของพารามิเตอร์ที่เรายังไม่ทราบค่าด้วย statistic ตัวหนึ่งซึ่งเป็น–ฟังก์ชันของข้อมูล ตัวอย่างที่ถูกเลือกขึ้นมา พารามิเตอร์ที่กล่าวนี้อาจเป็นค่ารวมทั้งหมด เช่นจำนวนประชากรทั้งประเทศ หรือเป็นค่าเฉลี่ย เช่น รายได้เฉลี่ยต่อบุคคลต่อปี หรือเป็นอัตราส่วน เช่น อัตราส่วนของรายได้ต่อรายจ่ายของครอบครัวในจังหวัดพระนครและธนบุรี

อีกปัญหาหนึ่งที่เราสนใจก็คือ ปัญหาทางด้านการทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ที่เรามีอยู่ พารามิเตอร์ดังกล่าวนั้นอาจได้จากประสบการณ์หรือจากการทดลองก็ตาม เราสนใจที่จะทดสอบในแง่ที่ว่า สมมติฐานที่เรายอมรับกันมานั้นมีความถูกต้องหรือไม่เพียงใด หรือในบางครั้งอาจมุ่งหมายที่จะทดสอบแบบของ–ฟังก์ชันของการกระจายของความน่าจะเป็นของข้อมูลตัวอย่างเป็นต้น

๒.๒ ความหมายของคำที่จะใช้

คำที่นับว่ามีความสำคัญมากในบทนี้อาจกล่าวว่ามีอยู่สองคำคือ

Estimator คือสูตร หรืออยู่ในรูปของฟังก์ชั่น

ตัวอย่างที่ ๒.๒.๑ สมมติว่าเรามีประชากรอยู่ชุดหนึ่งประกอบด้วย

$X_1, X_2, \dots, X_N$

$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$ ซึ่งเป็นพารามิเตอร์ที่เราไม่สามารถทราบค่าได้

$x_1, x_2, \dots, x_n$ เป็นตัวอย่างขนาด n หน่วยที่ถูกเลือกขึ้นมา

ถ้า $f(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$

แล้ว $f(x)$ ก็คือ estimator สำหรับพารามิเตอร์ μ

หรือถ้าเราจะทำการจัดเรียงข้อมูลตัวอย่างชุดนั้นเสียใหม่ตามขนาดของค่าจากสูงสุดไปยังค่าต่ำสุด สมมติให้เป็น $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ โดยที่

$x_{(1)} < x_{(2)} < x_{(3)} < \dots < x_{(n)}$

ถ้า $g(x) = \frac{x_{(1)} + x_{(n)}}{2}$ แล้ว $g(x)$ ก็คือ estimator สำหรับพารามิเตอร์ μ ด้วย

เป็นที่น่าสังเกตว่าในประชากรชุดหนึ่ง ๆ เราสามารถมี estimator ได้หลายตัวด้วยกัน ดังนั้นเป็นปัญหาที่นักสถิติจะต้องพิจารณาว่า estimator ใดจะเป็นตัวที่ดีที่สุด ซึ่งจะได้กล่าวถึงการเปรียบเทียบ estimator ที่ดีต่อไป

Estimate คือค่าตัวเลขจำนวนหนึ่ง ได้จากการแทนค่าตัวเลขจริง ๆ ที่ได้จากข้อมูลตัวอย่างลงในสูตรหรือฟังก์ชั่นที่เป็น estimator

เช่นจากตัวอย่างที่ ๒.๒.๑ ให้ตัวอย่างมีขนาด ๕ หน่วยประกอบด้วย

๑๒๙, ๑๓๖, ๑๔๑, ๑๔๔, ๑๕๑.

ถ้าใช้สูตร $f(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ จะได้ค่าประมาณเป็น

$$\frac{๑๒๙ + ๑๓๖ + ๑๔๑ + ๑๔๔ + ๑๕๑}{๕} = ๑๔๐.๒$$

ถ้าใช้สูตร $g(x) = \frac{x_{(1)} + x_{(n)}}{2}$ จะได้ค่าประมาณเป็น $\frac{๑๒๙ + ๑๕๑}{๒}$ = ๑๔๐

๒.๓ การประมาณค่า (Estimation)

การประมาณค่าคือการใช้ค่า Statistic มาประมาณค่าของพารามิเตอร์[๑] เมื่อใดก็ตามที่เราสามารถหาค่าจากข้อมูลตัวอย่างมาแทนใน Statistic ซึ่งเป็นฟังก์ชันของข้อมูลตัวอย่างได้ เราย่อมหาค่าประมาณสำหรับพารามิเตอร์นั้น ๆ ได้ ค่าประมาณ (estimate) แยกได้เป็นสองประเภทคือค่าเดี่ยว เช่น ค่าประมาณของค่าเฉลี่ยของอายุของนิสิตในจุฬาลงกรณมหาวิทยาลัยเป็น ๒๐ ปี ค่าประมาณของอายุการใช้งานของหลอดไฟฟ้าชนิดหนึ่งเป็น ๒๔๐.๗ ชั่วโมง เป็นต้น ค่าประมาณอีกประเภทหนึ่งเป็นค่าในช่วงคือค่าประมาณจะอยู่ในระหว่างค่าสองค่า เช่น ค่าประมาณของอุปสงค์ที่มีต่อปูนซิเมนต์ในปีหน้าจะอยู่ระหว่าง ๑๒๓.๔ ถึง ๑๔๓.๖ ล้านตัน เครื่องจักรชนิดใหม่จะให้ผลิตภัณฑ์ที่เสียเป็นค่าประมาณ ๔ ถึง ๔.๙ เปอร์เซ็นต์ของผลิตภัณฑ์ทั้งหมด เป็นต้น

๒.๔ Unbiased Estimates

ถ้า $\hat{\theta}(x)$ เป็น estimator สำหรับพารามิเตอร์ θ เราจะกล่าวว่า $\hat{\theta}(x)$ เป็น unbiased estimate ของ θ ถ้า

---

[๑] Paul G. Hoel, Introduction to Mathematical Statistics (3rd ed.; New York: John Wiley and Sons, Inc., 1962), p.96.

$$E[\hat{\theta}(x)] = \theta$$ [2]

ตัวอย่างที่ ๒.๔.๑ ประชากรชุดหนึ่งประกอบด้วย $X_1, X_2, \ldots, X_N$ และค่าตัวกลางเลขคณิต (mean) $= \mu = \frac{1}{N}\sum_{i=1}^{N} X_i$

ถ้าเราสุ่มตัวอย่างขนาด n หน่วยออกมาชุดหนึ่งเป็น $X_1, X_2, \ldots, X_n$ โดยที่ $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ แล้ว เราสามารถกล่าวได้ว่า $\bar{X}$ เป็น Unbiased estimate สำหรับ μ เพราะ

$$\begin{aligned} E(\bar{X}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= E\left[\frac{1}{n}(X_1+X_2+\cdots+X_n)\right] \\ &= \frac{1}{n}\left[E(X_1+X_2+\cdots+X_n)\right] \\ &= \frac{1}{n}\left[E(X_1)+E(X_2)+\cdots+E(X_n)\right] \\ &= \frac{1}{n}\left[\mu+\mu+\cdots+\mu\right] \\ &= \frac{n}{n}\mu = \mu \end{aligned}$$

2.5 Minimum Variance Unbiased Estimator

$\hat{\theta}(X)$ จะเป็น minimum variance unbiased estimator สำหรับ $\theta(X)$ ถ้า $\mathrm{Var}\ \hat{\theta}(X) \leq \mathrm{Var}\ \hat{\theta}^*(X)$ โดยที่ $\hat{\theta}^*(X)$ เป็น unbiased estimator ใด ๆ สำหรับ $\theta(X)$ [3]

---

[2] *Ibid.*, p.222.

[3] Alexander M. Mood and Franklin A. Graybill, *Introduction to the Theory of Statistics* (2nd ed.; New York: Mc Graw-Hill Book Company, Inc., 1963), p. 175.

<u>ตัวอย่างที่ ๒.๕.๑</u> ถ้าเราต้องการประมาณค่า $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

ก. ถ้าเราทำแบบ simple random sampling

$$\bar{y} = \frac{1}{n}\sum^{n} y_i \qquad (๒.๕.๑)$$

ซึ่ง $E(\bar{y}) = \bar{Y}$

และ $$\text{Var}(\bar{y})_{ran} = \frac{S^2}{n} \qquad (๒.๕.๒)$$

ข. ถ้าเราทำแบบ proportional stratified random sampling

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{L} n_h \bar{y}_h \qquad (๒.๕.๓)$$

ซึ่ง $E(\bar{y}) = \bar{Y}$

แต่ $$\begin{aligned}\text{Var}(\bar{y})_{prop} &= \Sigma N_h S_h^2 / nN \\ &= \frac{S^2}{n} - \Sigma N_h(\bar{Y}_h - \bar{Y})^2 / nN \\ &= \text{Var}(\bar{y})_{ran} - \Sigma N_h(\bar{Y}_h - \bar{Y})^2 / nN \qquad (๒.๕.๔)\end{aligned}$$

ดังจะเห็นได้ว่าในการประมาณค่าของ $\bar{Y}$ นั้นจะทำแบบ simple random sampling หรือ proportional stratified random sampling ก็จะได้ unbiased estimator เช่นเดียวกัน แต่จาก (๒.๕.๔) แสดงให้เห็นได้ว่า ถ้าเราประมาณค่า $\bar{Y}$ ด้วย (๒.๕.๓) จะให้ค่าความแปรปรวนน้อยกว่าการใช้ (๒.๕.๑)

เป็นที่น่าสังเกตว่าในการประมาณค่าทางสถิตินั้น เราสามารถจะมีตัวประมาณค่า (estimator) ได้มากกว่าหนึ่งตัวเสมอไป และในระหว่างบรรดาตัวประมาณค่าทั้งหลายนี้การจะตัดสินว่าตัวใดจะดีอย่างไรนั้นโดยปกติก็จะดูจากคุณสมบัติด้านต่าง ๆ ที่นับว่าสำคัญมากคือเปรียบเทียบด้วย ๒.๔ และ ๒.๕ นอกจากนั้นก็เปรียบเทียบด้วยคุณสมบัติอื่น ๆ ซึ่งเป็นคุณสมบัติทางด้านคณิตศาสตร์ เช่น ด้าน consistency ด้าน asymptotic และด้านความสะดวกในการคำนวณ (ease of computation) เป็นต้น

๒.๖ การทดสอบสมมติฐาน

ในการแก้ปัญหาทางสถิตินั้น นอกจากเราจะสนใจในการประมาณค่าของพารามิเตอร์แล้ว เรายังอาจสนใจที่จะทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ที่เรามีอยู่ พารามิเตอร์ดังกล่าวเราอาจได้มาจากประสพการณ์หรือจากการทดลองก็ตาม เราสนใจทดสอบสมมติฐานของพารามิเตอร์ในด้านที่ว่าจะมีการเปลี่ยนแปลงไปหรือไม่ หรืออาจทดสอบว่าสมมติฐานที่ยอมรับกันมานั้นถูกต้องหรือไม่เพียงใด หรือในบางครั้งอาจมุ่งหมายที่จะทดสอบแบบของฟังก์ชั่นของการกระจายของความน่าจะเป็นก็ได้

ตัวอย่าง กองส่งเสริมการเกษตรของกระทรวงเกษตร ต้องการจะสนองนโยบายการเพิ่มรายได้ประชาชาติของรัฐบาล โดยพยายามเพิ่มรายได้ในส่วนของผลิตผลการเกษตรให้สูงขึ้น และสนใจที่จะส่งเสริมการปลูกข้าวฟ่างนอกเหนือไปจากการส่งเสริมการปลูกข้าวจ้าวที่ได้ทำอยู่แล้ว จึงต้องการเปรียบเทียบผลตอบต่อไร่ของข้าวจ้าวและข้าวฟ่างดู

เช่นนี้จะเห็นได้ว่า กองส่งเสริมการเกษตรฯ จะต้องยุ่งเกี่ยวกับประชากรถึงสองชุดคือข้าวจ้าวกับข้าวฟ่าง สมมติว่าจากผลการทำนาข้าวจ้าวแต่ละปีที่ผ่านมา ผลตอบแทนต่อไร่โดยเฉลี่ยประมาณปีละ ๑,๔๐๐ บาท ปัญหาก็มีว่าถ้าปลูกข้าวฟ่างจะได้ผลตอบแทนต่อไร่มากกว่าหรือน้อยกว่าการปลูกข้าวจ้าว เพื่อขจัดข้อข้องใจนี้ เราก็ตั้งข้อสมมติฐานขึ้นว่า การเพาะปลูกพืชชนิดใดจะให้ผลตอบแทนต่อไร่สูงกว่า จากนั้นโดยการอาศัยการเลือกตัวอย่างจากการปลูกข้าวจ้าวและจากการทดลองปลูกข้าวฟ่าง เราก็สามารถจะเห็นพ้องหรือปฏิเสธข้อสมมติฐานของเราได้

ในทางปฏิบัติเราทดลองปลูกข้าวฟ่างแต่เพียงอย่างเดียว เพราะเรามีข้อมูลในด้านการเพาะปลูกข้าวจ้าวอยู่แล้ว สมมติว่าผลเฉลี่ยของการปลูกข้าวฟ่างได้ผลตอบต่อไร่เป็น ๑,๕๕๐ บาทต่อปี โดยผลสรุปอย่างหยาบ เราก็สามารถสรุปได้ว่า การปลูกข้าวฟ่างจะให้ผลตอบแทนต่อไร่สูงกว่าการปลูก

ข้าวจ้าว แต่ถ้าเราจะหันมาพิจารณาค่าเบี่ยงเบนมาตรฐานประกอบด้วยแล้วสมมติว่าคำนวณออกมาได้ค่าจากกระทงนา ตัวอย่างเป็น ๑๒๕ ($6/\sqrt{n}$ = ๑๒๕ เมื่อ n เป็นขนาดของจำนวนกระทงนาตัวอย่าง) ด้วยการใช้ ๙๕ % confidence interval แล้ว สำหรับการปลูกข้าวฟ่างจะมี interval อยู่ระหว่าง ๑,๓๐๐ ถึง ๑,๘๐๐ บาทต่อไร่ ถ้าสมมติฐานของเรากล่าวว่า "การปลูกข้าวฟ่างให้ผลตอบแทนสูงกว่าการปลูกข้าวจ้าว" แล้ว เราก็จะต้องปฏิเสธสมมติฐานดังกล่าว ด้วยเหตุผลที่ว่าผลตอบแทนต่อไร่ปีละ ๑,๕๕๐ บาทนั้นมีโอกาสสูงที่จะมาจากประชากรที่มีค่าเฉลี่ยเป็น ๑,๔๐๐ บาท แต่ในทางตรงกันข้าม ถ้าค่าเบี่ยงเบนมาตรฐานเป็น ๒๕ ($6/\sqrt{n}$ = ๒๕)เราจะต้องยอมรับข้อสมมติฐานที่ว่าการปลูกข้าวฟ่างให้ผลตอบแทนสูงกว่าการปลูกข้าวจ้าว