หลักการประมาณด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimation) และ โมเมนต์ (Method of Moments)

รศ.ดร. วีระชาติ กิเลนทอง
มหาวิทยาลัยหอการค้าไทย

บทนำ

- หัวข้อนี้จะนำเสนอหลักการพื้นฐานของการประมาณค่า (estimation) โดยพิจารณาวิธีการประมาณค่าอีก 2 วิธี
  - วิธีการประมาณค่าด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimation หรือเรียกสั้นๆ ว่า MLE)
  - วิธีการประมาณค่าด้วยโมเมนต์ (Method of Moments หรือเรียกสั้นๆ ว่า MM)

นิยาม: ฟังก์ชันความเป็นไปได้ (likelihood function)

- หลักการสำคัญของการประมาณค่าด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimation หรือเรียกสั้นๆ ว่า MLE) คือ ข้อมูลที่ได้จากการสังเกตเกิดจากการสุ่มเลือกมาจากประชากรที่มีการแจกแจงที่ **ขึ้นอยู่กับค่าพารามิเตอร์เฉพาะค่าหนึ่ง** ซึ่งควรจะเป็นค่าที่ทำให้ความเป็นไปได้ (likelihood) ที่จะสุ่มเลือกได้ตัวอย่างดังกล่าวมีค่าสูงสุด

Definition

ฟังก์ชันความเป็นไปได้ (likelihood function) คือ **ฟังก์ชันความหนาแน่นของความน่าจะเป็นร่วม** (joint p.d.f.) หรือ **ฟังก์ชันความน่าจะเป็นร่วม** (joint p.f.) $f(\boldsymbol{x}|\theta)$ โดยที่ $\boldsymbol{x} = (x_1, \ldots, x_n)$ คือ **ค่าที่เกิดขึ้นจริงที่ได้จากข้อมูล**

นิยาม: ฟังก์ชันความเป็นไปได้ (likelihood function)

- รูปแบบของฟังก์ชันความเป็นไปได้ (likelihood function) ขึ้นอยู่กับ **ข้อมูลที่ได้จากการสังเกต**
  - หากตัวแปรที่เกี่ยวข้องเป็นแบบไม่ต่อเนื่อง (discrete) ฟังก์ชันความเป็นไปได้ (likelihood function) ก็จะอยู่ในรูปแบบฟังก์ชันความน่าจะเป็นร่วม (joint p.f.) ถึงแม้ว่าตัวแปรสุ่มพื้นฐานที่ระบุในแบบจำลองจะเป็นแบบต่อเนื่องก็ตาม
- อีกด้านหนึ่ง ฟังก์ชันความเป็นไปได้ (likelihood function) ก็เป็น **ฟังก์ชันของพารามิเตอร์ที่ต้องการทราบค่า** θ โดยมองว่าค่าของ $\boldsymbol{x} = (x_1, \ldots, x_n)$ เป็นเพียงค่าคงที่ซึ่ง **ทราบค่าแล้วจากข้อมูลที่มีอยู่**
- ตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) หมายถึง **การแก้ปัญหาการหาค่าสูงสุดซึ่งมีฟังก์ชันความเป็นไปได้ (likelihood function) เป็นฟังก์ชันจุดประสงค์ (objective function)** และพารามิเตอร์ที่ต้องการทราบค่า θ เป็นตัวแปรที่ต้องเลือก

นิยาม: ตัวประมาณด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimation)

Definition

พิจารณาข้อมูลที่สังเกตได้ (observed data) $\mathbf{x}$ ชุดหนึ่ง และ $\hat{\theta}_{MLE}(\mathbf{x})$ เป็นคำตอบของปัญหาการหาค่าสูงสุด (maximization problem) ต่อไปนี้

$$f\left(\mathbf{x}|\hat{\theta}_{MLE}(\mathbf{x})\right) = \max_{\theta} f(\mathbf{x}|\theta) \tag{1}$$

หรือเขียนอีกรูปแบบหนึ่งได้เป็น

$$\hat{\theta}_{MLE}(\mathbf{x}) = \operatorname*{argmax}_{\theta} f(\mathbf{x}|\theta) \tag{2}$$

แล้ว **ฟังก์ชัน** $\hat{\theta}_{MLE}(\mathbf{X})$ คือ **ตัวประมาณค่า** ด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimator หรือเรียกสั้นๆ ว่า MLE) และ $\hat{\theta}_{MLE}(\mathbf{x})$ คือ **ค่าประมาณ**ด้วยความเป็นไปได้สูงสุดของ θ เมื่อ**ข้อมูลที่ใช้ในการประมาณค่า**คือ $\mathbf{x}$

การประมาณค่าด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimation)

- ฟังก์ชันความเป็นไปได้ (likelihood function) มักอยู่ในรูปของ **ผลคูณของฟังก์ชันความหนาแน่นของความน่าจะเป็น (p.d.f.) หรือฟังก์ชันความน่าจะเป็น (p.f.)**
- เพื่อความสะดวกจึงมักจะ **แปลงฟังก์ชันจุดประสงค์ (objective function) ให้อยู่ในรูปของล็อกการิธึม** แทน ซึ่งจะอยู่ในรูปของผลบวกซึ่งจัดการได้ง่ายกว่ามาก และที่สำคัญยังได้คำตอบเท่าเดิม
- การแปลงค่าฟังก์ชันจุดประสงค์โดยใช้ **ฟังก์ชันที่เพิ่มขึ้นทางเดียว (monotonically increasing)** จะ **ไม่** ทำให้คำตอบของปัญหาการหาค่าสูงสุดเปลี่ยนไป

ตัวอย่าง: การประมาณค่าด้วยการแจกแจงแบบเบอร์นูลลี (Bernoulli distribution)

Example

พิจารณาตัวอย่าง $X_1, \ldots, X_n$ โดยที่ X_i มีการแจกแจงแบบเบอร์นูลลี (Bernoulli distribution) สำหรับพารามิเตอร์ θ ซึ่งเป็นสิ่งที่ต้องการประมาณค่า และกำหนดให้ $\boldsymbol{x} = (x_1, \ldots, x_n)$ คือค่าที่เกิดขึ้นจริงที่ได้จากข้อมูล

- ฟังก์ชันความเป็นไปได้ (likelihood function) ในกรณีนี้เท่ากับ

$$\mathscr{L} = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \tag{3}$$

- ฟังก์ชันล็อกการิธึมของความเป็นไปได้ (log-likelihood function) เท่ากับ

$$\begin{aligned} \log \mathscr{L} &= \sum_{i=1}^{n} \left[x_i \log \theta + (1-x_i) \log (1-\theta) \right] \\ &= \left(\sum_{i=1}^{n} x_i \right) \log \theta + \left(n - \sum_{i=1}^{n} x_i \right) \log (1-\theta) \end{aligned}$$

สรุป: การประมาณค่าด้วยการแจกแจงแบบเบอร์นูลลี (Bernoulli distribution)

Example

- ปัญหาการหาค่าสูงสุดในกรณีคือ

$$\max_{\theta} \left(\sum_{i=1}^{n} x_i \right) \log \theta + \left(n - \sum_{i=1}^{n} x_i \right) \log (1 - \theta)$$

- เราสามารถหาคำตอบได้โดยการกำหนดให้ค่าอนุพันธ์อันดับที่หนึ่งของฟังก์ชันจุดประสงค์ (objective function) หรือฟังก์ชันล็อกการิธึมของความเป็นไปได้ (log-likelihood function) เท่ากับศูนย์ หรือที่เรียกว่า เงื่อนไขอันดับที่หนึ่ง (first-order conditions)

$$\left.\frac{\partial \log \mathscr{L}}{\partial \theta}\right|_{\hat{\theta}_{MLE}(x)} = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i \right) \frac{1}{\hat{\theta}_{MLE}(x)} - \left(n - \sum_{i=1}^{n} x_i \right) \frac{1}{1 - \hat{\theta}_{MLE}(x)} = 0$$

$$\Rightarrow \hat{\theta}_{MLE}(x) = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}_n$$

- ตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE): $\hat{\theta}_{MLE}(X) = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_n$
- ตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) $\hat{\theta}_{MLE}(X)$ เป็น **ฟังก์ชันของตัวอย่าง** X แต่หลังจากนี้ต่อไป จะขอไม่เขียนส่วนนี้ ทั้งนี้เพื่อความกระชับ โดยจะเขียนสั้นๆ เป็น $\hat{\theta}_{MLE}$ แต่ขอให้เข้าใจว่ามันคือ $\hat{\theta}_{MLE}(X)$

ตัวอย่าง: การประมาณค่าด้วยการแจกแจงแบบปกติ (normal distribution)

Example

สมมุติให้ $X_1, \ldots, X_n$ เป็นกลุ่มตัวอย่างสุ่มที่เลือกมาการแจกแจงแบบปกติ (normal distribution) ที่มีค่าคาดหมาย μ และค่าความแปรปรวน σ^2 โดยทั้งสองเป็นพารามิเตอร์ที่ต้องการทราบค่า และกำหนดให้ $\mathbf{x} = (x_1, \ldots, x_n)$ คือค่าที่เกิดขึ้นจริงที่ได้จากข้อมูล

- ฟังก์ชันความเป็นไปได้ (likelihood function) ในกรณีนี้เท่ากับ

$$\mathscr{L} = \prod_{i=1}^{n} \phi \left(\frac{x_i - \mu}{\sigma} \middle| \mu, \sigma^2 \right) \tag{4}$$

- ฟังก์ชันล็อกการิธึมของความเป็นไปได้ (log-likelihood function) เท่ากับ

$$\log \mathscr{L} = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{5}$$

- ปัญหาการหาค่าสูงสุดในกรณีคือ

$$\max_{\mu, \sigma^2} -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

สรุป: การประมาณค่าด้วยการแจกแจงแบบปกติ (normal distribution)

Example

- เงื่อนไขอันดับที่หนึ่ง (first-order conditions) ของปัญหานี้เท่ากับ

$$\left.\frac{\partial \log \mathscr{L}}{\partial \mu}\right|_{\hat{\mu}_{MLE},\hat{\sigma}^2_{MLE}} = 0 \Rightarrow \frac{1}{\hat{\sigma}^2_{MLE}} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE}) = 0 \tag{6}$$

$$\left.\frac{\partial \log \mathscr{L}}{\partial \sigma^2}\right|_{\hat{\mu}_{MLE},\hat{\sigma}^2_{MLE}} = 0 \Rightarrow -\frac{n}{2\hat{\sigma}^2_{MLE}} + \frac{1}{2\left(\hat{\sigma}^2_{MLE}\right)^2} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2 = 0 \tag{7}$$

- ระบบสมการข้างต้นมีสองตัวแปร และสามารถหาคำตอบได้เป็น

$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}_n \quad , \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}(x_i - \bar{x}_n)^2 \tag{8}$$

- เรามักเรียกตัวประมาณค่าสำหรับค่าคาดหมาย $\bar{X}_n$ ว่า **ค่าคาดหมายจากตัวอย่าง (sample mean)**
- ตัวประมาณค่าสำหรับค่าความแปรปรวน $\frac{1}{n}(X_i - \bar{X}_n)^2$ ว่า **ค่าความแปรปรวนจากตัวอย่าง (sample variance)** ซึ่งแตกต่างจาก $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ ตรงที่ส่วนตัวหารที่ $\hat{\sigma}^2_{MLE}$ นั้นหารด้วย n แต่ S_n^2 หารด้วย $n-1$
- ตัวประมาณค่าทั้งสองของความแปรปรวนนี้มี **คุณสมบัติลู่เข้าเชิงความน่าจะเป็นที่เหมือนกัน** เพราะ n และ $n-1$ นั้นมีพฤติกรรมที่ลิมิตเหมือนกัน แต่มี **คุณสมบัติความไม่เบี่ยงเบน (unbiasness) ต่างกัน** อภิปรายในรายละเอียดในหัวข้อถัดไป

ตัวอย่าง: การประมาณค่าด้วยการแจกแจงแบบเอกรูป (uniform distribution)

Example

สมมุติให้ $X_1, \ldots, X_n$ เป็นกลุ่มตัวอย่างสุ่มที่เลือกมาการแจกแจงแบบเอกรูป (uniform distribution) ในช่วง $[0, c]$ โดยในที่นี้ c คือพารามิเตอร์ที่ต้องการทราบค่า และกำหนดให้ $\boldsymbol{x} = (x_1, \ldots, x_n)$ คือค่าที่เกิดขึ้นจริงที่ได้จากข้อมูล

- ฟังก์ชันความเป็นไปได้ (likelihood function) ในกรณีนี้เท่ากับ

$$\mathscr{L} = \prod_{i=1}^{n} \frac{1}{c} = \frac{1}{c^n}, \text{ สำหรับ } 0 \leq x_i \leq c \text{ สำหรับทุก } i = 1, \ldots, n \tag{9}$$

- ส่วนกรณีอื่นๆ $\mathscr{L} = 0$ ปัญหาการหาค่าสูงสุดในกรณีคือ

$$\max_{c} \frac{1}{c^n}, \quad \text{โดยที่ } 0 \leq x_i \leq c \text{ สำหรับทุก } i = 1, \ldots, n$$

สรุป: การประมาณค่าด้วยการแจกแจงแบบเอกรูป (uniform distribution)

- ปัญหานี้เป็นปัญหาการหาค่าสูงสุด **แบบมีข้อจำกัด** (constrained optimization problem)
  - **หากไม่มีข้อจำกัด** คำตอบของ c ที่ทำให้ค่าฟังก์ชันความเป็นไปได้ (likelihood function) มีค่าสูงสุดคือ $c = 0$
  - แต่จากข้อจำกัด $0 \leq x_i \leq c$ ทำให้ค่า c จะต้องไม่น้อยกว่าค่า x_i แต่ละค่า ซึ่งทำให้ $c = 0$ เป็นไปไม่ได้ (ยกเว้น $X_i = 0$ ทุกค่า ซึ่งเป็นกรณีที่ไม่น่าสนใจ)
  - คำตอบของปัญหานี้จึงถูกกำหนดโดยข้อจำกัดเป็นหลัก นั่นคือ $\hat{c}_{MLE} \geq x_i$ สำหรับทุก $i = 1, \ldots, n$
  - แต่ในขณะเดียวกันฟังก์ชันวัตถุประสงค์ก็ต้องการให้ค่า $\hat{c}_{MLE}$ มีค่าต่ำที่สุดเท่าที่จะเป็นไปได้ ดังนั้น **คำตอบ** ที่ได้คือ $\hat{c}_{MLE} = \max(x_1, \ldots, x_n)$
- บทเรียนทางคณิตศาสตร์อันหนึ่งในตัวอย่างนี้คือ บางครั้งเราอาจจะ **ไม่สามารถแก้ปัญหา**การหาค่าสูงสุดแบบมีข้อจำกัด (constrained optimization problem) ได้โดยใช้ **เงื่อนไขอันดับที่หนึ่ง (first-order conditions)** โดยตรงเหมือนกับตัวอย่างที่ผ่านมา
- ทั้งนี้เพราะบางครั้ง **คำตอบที่ได้อาจจะอยู่บนขอบ (boundary) ของขอบเขตที่เป็นไปได้** ในขณะที่ เงื่อนไขอันดับที่หนึ่ง (first-order conditions) ใช้ได้กับกรณีที่คำตอบอยู่ภายในขอบเขตที่เป็นไปได้ (interior solution)

ทฤษฎี: คุณสมบัติไม่แปรเปลี่ยน (invariance property) ของตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

- ตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) มีคุณสมบัติอย่างหนึ่งที่สำคัญที่ช่วยให้สามารถประยุกต์ใช้ได้อย่างกว้างขวาง โดยเรามักเรียกคุณสมบัตินี้ว่า **คุณสมบัติไม่แปรเปลี่ยน (invariance property)** ซึ่งมีความหมายดังแสดงในทฤษฎีบทต่อไปนี้

Theorem

ถ้า $\hat{\theta}_{MLE}$ *คือตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) ของ* θ *และ* g *เป็นฟังก์ชันที่กำหนดโดย* $\gamma = g(\theta)$ *แล้ว* $g\left(\hat{\theta}_{MLE}\right)$ *เป็นตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) ของ* γ

การพิสูจน์: คุณสมบัติไม่แปรเปลี่ยน (invariance property) ของตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

Proof.

- เพื่อความสะดวก ขอนำเสนอ **การพิสูจน์ในกรณีที่ g เป็นฟังก์ชันหนึ่งต่อหนึ่ง (one-on-one) เท่านั้น** โดยเริ่มจาก การที่ $\hat{\theta}_{MLE}$ เป็นตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) ของ θ หมายความว่า

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} f_i \left(x_i | \theta \right) \tag{10}$$

- การที่ g เป็นฟังก์ชันหนึ่งต่อหนึ่ง (one-on-one) หมายความว่า เราสามารถใช้ฟังก์ชันส่วนกลับได้เป็น $\theta = g^{-1}(\gamma)$ ดังนั้น เราสามารถเขียนฟังก์ชันความเป็นไปได้ (likelihood function) ในรูปฟังก์ชันของ γ ได้เป็น

$$\mathscr{L}\left(\gamma|\mathbf{x}\right) = \underset{\gamma}{\operatorname{argmax}} \prod_{i=1}^{n} f_i \left(x_i | g^{-1}(\gamma) \right) \tag{11}$$

- ดังนั้น เนื่องจากฟังก์ชันความเป็นไปได้ (likelihood function) ในสมการที่ 10 มีค่าสูงสุดเมื่อพารามิเตอร์ θ มีค่าเท่ากับ $\hat{\theta}_{MLE}$ ดังนั้น ฟังก์ชันความเป็นไปได้ (likelihood function) ในสมการที่ 11 จะมีค่าสูงสุด ถ้าพารามิเตอร์ γ มีค่าเท่ากับ $\hat{\gamma}_{MLE}$ ที่ทำให้ $\hat{\theta}_{MLE} = g^{-1}(\hat{\gamma}_{MLE})$ นั่นคือ $\hat{\gamma}_{MLE} = g\left(\hat{\theta}_{MLE}\right)$
- ส่วนในกรณีที่ฟังก์ชัน g ไม่ใช่ฟังก์ชันหนึ่งต่อหนึ่ง ก็สามารถพิสูจน์ได้ด้วยวิธีการที่คล้ายคลึงกัน แต่ต้องตระหนักว่า สิ่งที่ต้องการคือค่าที่ต้องการนำไปสู่ค่าสูงสุด ถึงแม้ว่าอาจจะมีค่าพารามิเตอร์ที่ทำให้ได้ค่าสูงสุดหลายค่า

$\square$

ตัวอย่าง: คุณสมบัติไม่แปรเปลี่ยน (invariance property) ของตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

Example

สมมุติให้ $X_1, \ldots, X_n$ เป็นกลุ่มตัวอย่างสุ่มที่เลือกมาการแจกแจงแบบเอกรูป (uniform distribution) ในช่วง $[0, c]$ เช่นเดียวกับตัวอย่างที่แล้วซึ่งพิสูจน์แล้วว่า **ตัวประมาณค่า** ด้วยความเป็นไปได้สูงสุด (MLE) ของ c มีค่าเท่ากับ $\hat{c}_{MLE} = \max(X_1, \ldots, X_n)$
เราสามารถประยุกต์ใช้คุณสมบัติไม่แปรเปลี่ยน (invariance property) เพื่อหา **ตัวประมาณค่า** ของความแปรปรวน (variance) ของ X ได้โดยเริ่มจากการที่ทราบว่า $Var[X] = \frac{c^2}{12}$ ในขณะเดียวกันเราก็ทราบว่า **ตัวประมาณค่า** $\hat{c}_{MLE} = \max(X_1, \ldots, X_n)$ ดังนั้น

$$\widehat{Var}[X] = \frac{\hat{c}_{MLE}^2}{12} = \frac{\max(X_1, \ldots, X_n)^2}{12}$$

ความคงเส้นคงวา (consistency) ของตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

- คุณสมบัติอันหนึ่งที่สำคัญของ **ตัวประมาณค่า** ด้วยความเป็นไปได้สูงสุด (MLE) คือ **ความคงเส้นคงวา (consistency)** นั่นคือ เมื่อ **กลุ่มตัวอย่างมีขนาดใหญ่มากพอ** แล้ว ตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE) จะมี **ค่าใกล้เคียงกับค่าที่แท้จริง**
- ทฤษฎีบทและการพิสูจน์ด้านล่างนี้ดัดแปลงมาจาก Hogg et al.(2005) โดยเริ่มจากการกำหนดเงื่อนไขปกติ (regularity conditions) ต่อไปนี้

  RC1 ฟังก์ชันความหนาแน่นของความน่าจะเป็น (p.d.f.) $f(x, \theta)$ **แตกต่างกันชัดเจน** หมายความว่า ถ้า $\theta \neq \theta'$ แล้ว $f(x, \theta) \neq f(x, \theta')$

  RC2 ฟังก์ชันความหนาแน่นของความน่าจะเป็น (p.d.f.) $f(x, \theta)$ มีส่วนค้ำจุน (support) S **เหมือนกันสำหรับทุกๆ ค่าพารามิเตอร์** θ

  RC3 ค่าที่แท้จริง θ_0 ของพารามิเตอร์ θ เป็น **จุดภายใน (interior point) ของเซ็ต** Θ

ทฤษฎี: ความคงเส้นคงวา (consistency) ของตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

Theorem

สมมุติว่าตัวแปรสุ่ม $X_1, \ldots, X_n$ *สอดคล้องกับเงื่อนไขปกติ (regularity conditions) RC1-RC3 โดยที่* θ_0 *เป็นค่าพารามิเตอร์ที่แท้จริง (true parameter) และ* $f(x|\theta)$ *สามารถหาค่าอนุพันธ์เทียบกับ* θ *ได้ แล้ว คำตอบของสมการความเป็นไปได้ (likelihood equation)* $\hat{\theta}$ *ซึ่งสอดคล้องกับสมการต่อไปนี้*

$$\frac{\partial \log \mathscr{L}\left(\hat{\theta}|X\right)}{\partial \theta} = 0 \tag{12}$$

จะมีความคงเส้นคงวา (consistent) นั่นคือ $\hat{\theta} \xrightarrow{p} \theta_0$

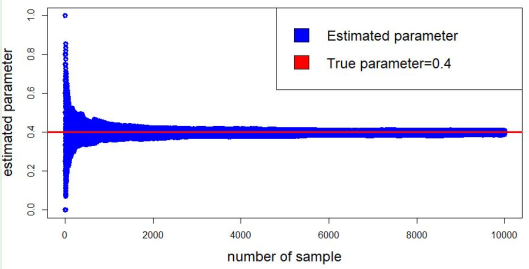
ตัวอย่าง: ความคงเส้นคงวา (consistency) ของตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

Example

พิจารณาตัวอย่าง $X_1, \ldots, X_n$ โดยที่ X_i มีการแจกแจงแบบเบอร์นูลลี (Bernoulli distribution) สำหรับพารามิเตอร์ θ ดังนั้นตัวประมาณค่าด้วยความเป็นไปได้สูงสุด (MLE)

$$\hat{\theta}_{MLE}(X) = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{x}_n \tag{13}$$

- เราสุ่มตัวอย่างจาก การแจกแจงแบบเบอร์นูลลี (Bernoulli distribution) โดยกำหนดพารามิเตอร์ $\theta = 0.4$ แล้ว นำข้อมูลจากการสุ่มนั้นมาประมาณพบว่า



บทนำ: การประมาณค่าด้วยโมเมนต์ (Method of Moments)

- **การประมาณค่าด้วยโมเมนต์ (method of moments หรือเรียกสั้นๆ ว่า MM)** เป็นวิธีประมาณค่าอย่างง่ายที่อาศัยหลักการพื้นฐานที่ว่า **ค่าคาดหมายจากตัวอย่าง (sample mean) $\bar{X}_n$ คือตัวประมาณค่า (estimator) ที่ดีสำหรับค่าคาดหมาย (population mean) $E[X]$** ซึ่งเป็นหลักการที่ประยุกต์ใช้อย่างกว้างขวาง แม้แต่ในกรณีของการประมาณค่าที่ซับซ้อน
- ข้อดีอีกอย่างหนึ่งของการประมาณค่าด้วยโมเมนต์ (MM) ก็คือการที่สามารถหาตัวประมาณค่า (estimator) ได้ค่อนข้างแน่นอน ซึ่งต่างจากวิธีการประมาณค่าด้วยความเป็นไปได้สูงสุด (Maximum Likelihood Estimation หรือเรียกสั้นๆ ว่า MLE) ที่อาจจะไม่สามารถหาตัวประมาณค่าได้ (estimator)

การประมาณค่าด้วยโมเมนต์ (Method of Moments)

- กำหนดให้ $X_1, \ldots, X_n$ เป็น **ตัวอย่างที่สุ่มมาจากประชากร** ที่มีฟังก์ชันความหนาแน่นของความน่าจะเป็น (p.d.f.) หรือฟังก์ชันความน่าจะเป็น (p.f.) $f(\mathbf{x}|\theta_1, \ldots, \theta_k)$
- วิธีการประมาณค่าด้วยโมเมนต์ (MM) ของพารามิเตอร์ $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ เริ่มได้ด้วยการ **กำหนดให้ k โมเมนต์จากตัวอย่าง (sample moments) มีค่าเท่ากับ k โมเมนต์จากประชากร (population moments)** ซึ่งเป็นฟังก์ชันของพารามิเตอร์ $\boldsymbol{\theta}$
- ผลลัพธ์จากขั้นตอนนี้คือ ระบบสมการของ k สมการที่มีตัวไม่ทราบค่า (unknown) ทั้งหมด k ตัว ขั้นตอนสุดท้ายคือการแก้ระบบสมการนี้ โดยที่ผลลัพธ์ที่ได้จะอยู่ในรูปของฟังก์ชัน $\theta_i(X_1, \ldots, X_n)$

โมเมนต์จากตัวอย่าง (sample moments)

- กำหนดให้ $M_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ แทน **โมเมนต์จากตัวอย่าง (sample moments)** อันดับที่ $j = 1, \ldots, k$ และ $\mu_j(\theta_1, \ldots, \theta_k) = E[X^j]$ แทน **โมเมนต์จากประชากร (population moments)** อันดับที่ $j = 1, \ldots, k$
- การประมาณค่าด้วยโมเมนต์ (MM) คือการแก้ระบบสมการต่อไปนี้

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1(\theta_1, \ldots, \theta_k) \tag{14}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu_2(\theta_1, \ldots, \theta_k) \tag{15}$$

$$\vdots \tag{16}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = \mu_k(\theta_1, \ldots, \theta_k) \tag{17}$$

- นี่คือระบบสมการที่มีทั้งหมด k สมการและมีตัวไม่ทราบค่าทั้งหมด k ตัว

ตัวอย่าง: การประมาณค่าด้วยโมเมนต์ (Method of Moments)

Example

สมมุติว่า เป็นตัวอย่างสุ่มที่มีมีการแจกแจงเหมือนกันและเป็นอิสระต่อกัน (i.i.d.) แบบปกติ $N\left(\mu, \sigma^2\right)$ นั่นคือ พารามิเตอร์ที่ต้องการทราบค่าในตัวอย่างนี้มีทั้งหมด 2 ตัวคือ $\theta_1 = \mu \; \theta_2 = \sigma^2$

- เราจำเป็นต้องใช้โมเมนต์ที่หนึ่งและโมเมนต์ที่สอง

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu \quad , \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = E\left[X^2\right] = \sigma^2 + \mu^2$$

- ตัวประมาณค่าด้วยโมเมนต์ (MM estimators) ในกรณีนี้เท่ากับ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \equiv \bar{X} \quad , \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

- ตัวประมาณค่าสำหรับค่าคาดหมายและค่าความแปรปรวนที่ใช้กันอยู่ทั่วไป อาจจะแตกต่างบ้างเล็กน้อยในกรณีของค่าความแปรปรวนที่หารด้วย n ไม่ใช่ $n-1$
- ความแตกต่างตรงนี้มีผลต่อความเบี่ยงเบน (bias) ของตัวประมาณค่าในทางทฤษฎี แต่ในทางปฏิบัติหาก n มีค่ามากพอแล้ว ค่าประมาณที่ได้จะมีค่าใกล้เคียงกันมาก

การประมาณค่าด้วยโมเมนต์ (Method of Moments)

- อย่างไรก็ตาม วิธีการประมาณค่าด้วยโมเมนต์ (MM) ก็มีจุดอ่อน ทั้งนี้ส่วนหนึ่งเป็นเพราะโมเมนต์เพียงไม่กี่อันอาจจะ**ไม่สามารถบอกคุณสมบัติของการแจกแจงได้ดีพอ** เพราะหากต้องการให้ครอบคลุมทั้งหมดต้องใช้โมเมนต์ทุกอัน
- ทฤษฎีบทต่อไปนี้สรุปว่า ตัวประมาณค่าด้วยโมเมนต์ (MM) มีความคงเส้นคงวา (consistent) นั่นคือ เมื่อจำนวนตัวอย่างมากพอแล้ว ค่าประมาณด้วยโมเมนต์ (MM) จะมีค่าใกล้เคียงกับค่าพารามิเตอร์ที่แท้จริง (true parameter)

ทฤษฎี: ความคงเส้นคงวา (consistency) ของการประมาณค่าด้วยโมเมนต์ (Method of Moments)

Theorem

สมมุติว่าตัวแปรสุ่ม $X_1, \ldots, X_n$ ที่เป็นอิสระต่อกันและมีการแจกแจงเหมือนกัน (i.i.d.) โดยมี $\boldsymbol{\theta}_0$ เป็นค่าพารามิเตอร์ที่แท้จริง (true parameter) ซึ่งมีทั้งหมด k ตัว และสมมุติว่า k โมเมนต์ของการแจกแจงหาค่าได้และมีค่าจำกัด แล้ว ตัวประมาณค่าด้วยโมเมนต์ (MM) จะมีความคงเส้นคงวา (consistent) นั่นคือ $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$

การพิสูจน์: ความคงเส้นคงวา (consistency) ของการประมาณค่าด้วยโมเมนต์ (Method of Moments)

Proof.

กฎว่าด้วยตัวอย่างขนาดใหญ่ (Law of Large Numbers) ทำให้สรุปได้ว่า สำหรับค่าจำนวนเต็ม $0 < j \leq k$

$$plim\ M_j = plim\ \frac{1}{n}\sum_{i=1}^{n} X_i^j = E\left[X_i^j\right]$$

ซึ่งมีค่าเท่ากับ $\mu_j\left(\boldsymbol{\theta}_0\right)$ ในขณะเดียวกัน เมื่อเราประยุกต์ใช้การลู่เข้าเชิงความน่าจะเป็นกับระบบสมการที่ (15)-(18) จะสามารถสรุปได้ว่า

$$plim\ M_j = plim\ \mu_j\left(\hat{\boldsymbol{\theta}}\right) = \mu_j\left(plim\ \hat{\boldsymbol{\theta}}\right)$$

ดังนั้นเราจึงสามารถสรุปได้ว่า

$$\mu_j\left(plim\ \hat{\boldsymbol{\theta}}\right) = \mu_j\left(\boldsymbol{\theta}_0\right) \Rightarrow plim\ \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$$

□

ตัวอย่างความคงเส้นคงวา (consistency) ของการประมาณค่าด้วยโมเมนต์ (Method of Moments)

Example

จากตัวอย่างที่แล้ว ถ้าเราสุ่มตัวอย่างที่มีมีการแจกแจงเหมือนกันและเป็นอิสระต่อกัน (i.i.d.) แบบปกติ $N\left(\mu, \sigma^2\right)$ ดังนั้น ตัวประมาณค่าด้วยโมเมนต์ (MM estimators) ในกรณีนี้เท่ากับ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \equiv \bar{X} \quad , \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- ถ้าตัวอย่างที่สุ่ม คือ การแจกแจงเหมือนกันและเป็นอิสระต่อกัน (i.i.d.) แบบปกติ ที่มี $\mu = 50$ และ $\sigma^2 = 50$