บทที่ 1
กฎการใช้กลุ่มตัวอย่างขนาดใหญ่
(LAW OF LARGE NUMBER)

1.1 มัชฉิมเลขคณิตของกลุ่มตัวอย่าง (Sample Mean)

ถ้า $(X_1, X_2, ..., X_n)$ เป็นกลุ่มตัวอย่าง (Sampled Random Variables)[1] ขนาด n ที่สุ่มมาจากกลุ่มประชากร X เราจะนิยามตัวแปรสุ่ม (Random Variable หรือ Vriate) $\bar{X}$ ซึ่งเป็นค่าเฉลี่ยของกลุ่มตัวอย่าง (Sample Mean) ดังกล่าวได้ดังนี้

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ให้ Sample Point คือ $(x_1, x_2, ..., x_n)$ เป็นค่าจริงที่ได้จากการวัดหรือการสังเกตเป็นค่าเฉพาะของตัวแปรสุ่ม $X_1, X_2, ..., X_n$ ตามลำดับ

ดังนั้น ค่าเฉพาะของตัวแปร $\bar{X}$ จึงปรากฏดังนี้

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

---

[1] โดยทั่วไปเราใช้อักษรตัวพิมพ์ใหญ่เช่น X, Y, Z, W, ... แทนตัวแปรสุ่ม และใช้อักษรตัวพิมพ์เล็กเช่น x, y, z, w, ... เป็นค่าเฉพาะที่ได้จากการสำรวจจริง หรือสังเกต เช่นค่าที่ได้จากการวัดหรือทดลองของตัวแปรสุ่ม (Random Variable) ดังกล่าว

ตัวแปรสุ่มหนึ่ง ๆ จะมีการแจกแจงของตัวเองเสมอ และขอให้ทำความเข้าใจไว้ในชั้นนี้ด้วยว่า โดยปกติแล้วในกลุ่มประชากรหนึ่ง ๆ นั้นมิได้หมายความว่าจะมีตัวแปรสุ่มของกลุ่มประชากรนั้นเพียงตัวเดียว แต่มิได้เป็นจำนวนมากมาย (Infinitely Many) ดังนั้น เมื่อดำเนินการสุ่มตัวอย่างกลุ่มตัวอย่างในชั้นต้นจึงเป็นกลุ่มของตัวแปรสุ่ม (Sampled Random Variables) จากกลุ่มประชากรนั้น จากนั้นจึงค่อยดำเนินการวัดหรือสังเกตเพื่อหาค่าเฉพาะของตัวแปรสุ่มเท่านั้น ค่าที่ได้จะเป็นตัวแทนของตัวแปรสุ่มแต่ละตัวเขียนแทนด้วยอักษรตัวพิมพ์เล็ก (ทางทฤษฎี) หรือค่าตัวเลข (ทางปฏิบัติ)

ดังนั้น ในทางการปฏิบัติเราจึงมักจะข้าม Concept นี้ไปเพราะถือว่าเป็นเรื่องที่ทุกคนเข้าใจกันดีอยู่แล้ว และพูดถึงค่าเฉลี่ยในรูปของ $\bar{x} = \sum_{i=1}^{n} x_i$ ในทันที

และเนื่องจาก $\bar{X}$ เป็นตัวแปรสุ่ม เราจึงสามารถหาค่าคาดหมายและความแปรปรวนของ $\bar{X}$ ได้ ซึ่งกระทำได้โดยง่าย กล่าวคือ

$$\begin{aligned} E(\bar{X}) &= E\,\frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\,E\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \end{aligned}$$

แต่เนื่องจากตัวแปร X_i มีการแจกแจงเช่นเดียวกันกับตัวแปรสุ่ม X (เพราะว่า X_i เป็นตัวแปรสุ่มที่สุ่มมาจากกลุ่มประชากร X) ดังนั้น

$$\begin{aligned} E(X_i) &= E(X) \\ &= \mu \end{aligned}$$

นั่นคือ

$$\begin{aligned} E(\bar{X}_i) &= E\,\frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\,E\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n}\,(n\mu) \\ &= \mu \end{aligned}$$

และ

$$\begin{aligned} V(\bar{X}) &= V\,\frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n^2}\,V\sum_{i=1}^{n} X_i \end{aligned}$$

แต่เนื่องจากตัวแปรสุ่ม X_i มีการแจกแจงเช่นเดียวกันกับตัวแปรสุ่ม X และ $(X_1, X_2, \ldots, X_n)$ เป็น Random Sample (ซึ่งหมายความว่า X_i เป็นอิสระต่อกัน)

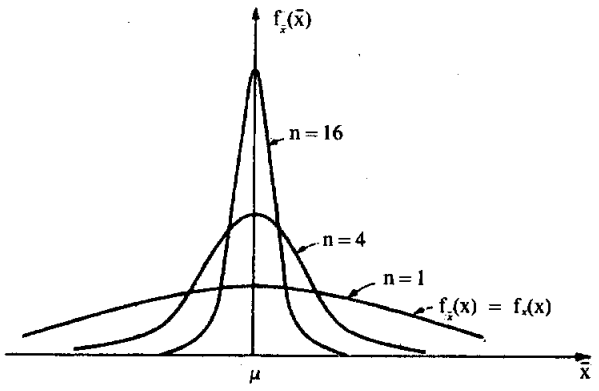
ดังนั้น $$V\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} V(X_i) \quad \text{(ไม่มี Covariance)}$$

และเนื่องจาก $V(X_i) = V(X) = \sigma^2$ ในทุกค่าของ i

ดังนั้น
$$\begin{aligned} V(\bar{X}) &= V\frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n^2} V\sum_{i=1}^{n} X_i \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(X) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n^2}(n\sigma^2) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

พิจารณา $V(\bar{X}) = \frac{\sigma^2}{n}$ จะพบว่าถ้ากลุ่มตัวอย่าง n มีขนาดใหญ่ (n มีค่ามาก) $\frac{\sigma^2}{n}$ จะมีค่าน้อยลง และทำให้น่าเชื่อได้ว่า ถ้ากลุ่มตัวอย่างมีขนาดใหญ่แล้ว $\bar{X}$ จะมีค่าใกล้กับค่าเฉลี่ย μ ด้วยโอกาสค่อนข้างสูง (หรือ $\bar{X}$ คาดหมายค่าของ μ ได้แม่นยำ)

เส้นโค้งแสดงการแจกแจงของ $\bar{X}$ สำหรับตัวอย่างขนาดต่าง ๆ ปรากฏดังภาพ



Distribution of $\bar{X}$ for n = 1, 4, 16.

เรื่องนี้จึงยังเป็นที่น่าสงสัยกันอยู่ อย่างไรก็ตามเรื่องนี้จะนำไปสู่การศึกษาเรื่อง กฎของการใช้กลุ่มตัวอย่างขนาดใหญ่ (Law of Large Number) ซึ่งจะใช้เป็นเครื่องมือตัดสินความสงสัยข้างต้น

ก่อนที่จะพุ่งความสนใจไปสู่กฎของการใช้กลุ่มตัวอย่างขนาดใหญ่โดยทันที ขอให้ศึกษาและพิจารณากฎเกณฑ์ หรือทฤษฎีที่คล้ายคลึงกันนี้คือ อสมการของเชพบิเชฟ ซึ่งมีลักษณะเป็นรูปทั่วไปและใช้เป็นหลักในการนำไปสู่การพิสูจน์กฎการใช้กลุ่มตัวอย่างขนาดใหญ่ และยังสามารถนำไปใช้ประโยชน์อย่างอื่นได้อีกมาก

1.2 อสมการของเชพบิเชฟ (Chebyshev's Inequality)

อสมการของเชพบิเชฟสามารถนำไปสู่การประยุกต์กับงานได้หลายลักษณะ โดยเฉพาะอย่างยิ่งสามารถใช้หาความน่าจะเป็นที่ค่าของตัวแปรสุ่มใด (Statistics) จะคลาดเคลื่อนไปจากค่าเฉลี่ย (ค่าคาดหมาย)[1] ทั้งนี้โดยมิได้มีข้อจำกัดว่าตัวแปรสุ่มจะต้องมีการแจกแจงเป็นรูปใดและไม่มีความจำเป็นจะต้องทราบการแจกแจง (Probability Distribution) ของตัวแปรสุ่มนั้นด้วย ขอเพียงให้ทราบค่าคาดหมายและค่าความแปรปรวนของตัวแปรสุ่มนั้นก็เป็นการเพียงพอแล้ว แต่ทั้งนี้มิได้หมายความว่าถ้าทราบการแจกแจง (Probability Distribution) ของตัวแปรสุ่มนั้นด้วยแล้วจะใช้อสมการ ของเชพบิเชฟไม่ได้ การที่ยิ่งทราบข้อสนเทศ (Information) เกี่ยวกับตัวแปรสุ่มที่สนใจมากเพียงใดจะยิ่งทำให้สามารถคำนวณหาค่าความน่าจะเป็นถูกต้องมากเพียงนั้น อนึ่งควรทราบว่าอสมการของเชพบิเชฟ สามารถให้ค่าความน่าจะเป็นได้เพียงค่าขีดจำกัดบน (Upper Bound) และขีดจำกัดล่าง (Lower Bound) เท่านั้น ไม่อาจให้ค่าที่แน่นอน (Exact Value)

---

[1] ค่าคาดหมายของตัวแปรสุ่ม ไม่จำเป็นต้องหมายถึงค่าเฉลี่ยเสมอไป ทั้งนี้สุดแต่ว่าตัวแปรสุ่มนั้นแทนสถานการณ์ใดเช่น ถ้าตัวแปรสุ่ม X แทนค่าที่เกิดจากการวัดเช่น ความสูง น้ำหนัก อายุ ราคา จำนวนต่าง ๆ ฯลฯ ค่า E(X) ก็หมายถึงค่าเฉลี่ย เช่น ความสูงเฉลี่ย น้ำหนักโดยเฉลี่ย อายุเฉลี่ย ราคาเฉลี่ย และอื่น ๆ ถ้าตัวแปรสุ่ม X แทนจำนวนครั้งของการทดลองจนประสบผลสำเร็จตามจำนวนครั้งที่ต้องการ หรือการรอคอยจนกระทั่งประสบผลสำเร็จสมหวัง (เช่น r ครั้งในกรณี Negative Binomial และครั้งแรกในกรณี Geometric Distribution) กรณีเช่นนี้ E(X) ควรเป็นค่าคาดหมาย (ไม่ควรเป็นค่าเฉลี่ย) ซึ่งหมายถึงจำนวนครั้งของการรอคอยซึ่งคาดว่าเมื่อทำการทดลองหรือรอคอยถึงจำนวนครั้งดังกล่าว ก็น่าจะประสบความสมหวังตามต้องการ

ได้ คำตอบที่ได้จาก อสมการของเชพบิเชฟจึงเป็นคำตอบกว้าง ๆ หรือให้ข้อเท็จจริงได้อย่างกว้าง ๆ เท่านั้น เช่น โอกาสที่รายได้ของครอบครัวใด ๆ จะคลาดไปจากรายได้เฉลี่ยของชาติตั้งแต่ 1,200 บาทขึ้นไป มีได้ไม่เกินกว่า 70% หรือ โอกาสที่ความต้องการสินค้าประเภทเสื้อผ้าของนาย ก. จะคลาดไปจากความต้องการโดยประมาณของผู้มีอาชีพเดียวกันไม่ถึง 3 ชิ้น มีได้ไม่ต่ำกว่า 90% ดังนี้เป็นต้น

อนึ่ง นอกจากประโยชน์ในการให้คำตอบได้อย่างกว้าง ๆ ดังกล่าวนี้แล้ว อสมการของเชพบิเชฟยังสามารถใช้เป็นเครื่องมือในการกำหนดขนาดของตัวอย่างเพื่อใช้ในการสำรวจอีกด้วย คำตอบหรือขนาดตัวอย่างที่ได้นี้จะออกมาในรูปของขีดจำกัดบนหรือขีดจำกัดล่างเช่นกัน

ทฤษฎี 1.1 ให้ตัวแปรสุ่ม X มีค่าคาดหมายและค่าความแปรปรวนเป็น $E(X) = \mu$ และ $V(X) = \sigma^2$ ตามลำดับ ถ้า ε เป็นเลขจำนวนเต็มบวกใด ๆ (ที่มีค่าน้อยมาก) แล้ว

ก. โอกาสที่ค่าของตัวแปรสุ่ม X จะคลาดไปจากค่าคาดหมาย E(X) ไปแม้แต่เพียงเล็กน้อย (ตั้งแต่ ε ขึ้นไป) จะมีได้ไม่เกินไปกว่า $\frac{V(X)}{\varepsilon^2}$ หรือนัยหนึ่ง

$$\Pr\{|X - E(X)| \geq \varepsilon\} \leq \frac{V(X)}{\varepsilon^2}$$

และ

ข. โอกาสที่ค่าของตัวแปรสุ่ม X จะคลาดไปจากค่าคลาดหมาย E(X) ไปแม้แต่เพียงเล็กน้อย (ไม่ถึง ε) จะมีค่าได้ไม่น้อยกว่า $1 - \frac{V(X)}{\varepsilon^2}$ หรืออีกนัยหนึ่ง

$$\Pr\{|X - E(X)| < \varepsilon\} \geq 1 - \frac{V(X)}{\varepsilon^2}$$

ข้อสังเกต ขอให้สังเกตว่าอสมการของเชพบิเชฟ พูดถึงความคลาดเคลื่อนในรูปของ Absolute Deviate งานที่จะใช้อสมการของเชพบิเชฟจึงควรมีลักษณะของความคลาดเคลื่อนที่เป็น Absolute Deviation เท่านั้น ถ้างานใดที่มีลักษณะนอกเหนือไปกว่านี้ให้ใช้หลักการอื่นอาจเป็น Central Limit Theorem หรือหลักการหาความน่าจะเป็นโดยวิธีอื่น ๆ ซึ่งจะกล่าวถึงในลำดับต่อไป

Absolute Deviation จะบอกให้ทราบเพียงว่าค่าสังเกตอาจมากกว่าหรือน้อยกว่าค่าคาดหมายไปจำนวนหนึ่งไม่ระบุว่ามากกว่าหรือน้อยกว่าอย่างชัดเจน แต่ระบุไว้เพียง "คลาดเคลื่อน" กันเท่านั้น

พิสูจน์[1]

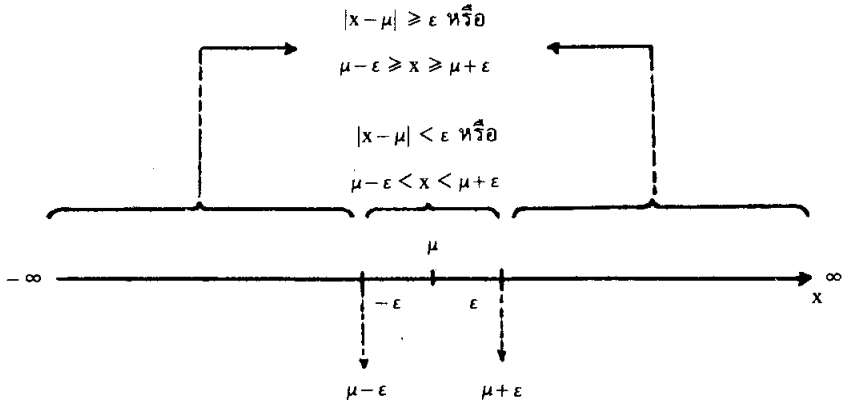
วิธีที่ 1 จากนิยามของค่าความแปรปรวนคือ

$$\sigma^2 = V(X) = E(X - E(X))^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

โดยที่ $\mu = E(X)$ และ $f(x)$ เป็น Probability Density Function (pdf) ของตัวแปรสุ่ม X นั่นคือ

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \\ &= \int_{|x-\mu|\geq\varepsilon} (x - \mu)^2 f(x)dx + \int_{|x-\mu|<\varepsilon} (x - \mu)^2 f(x)dx\end{aligned}$$

โดยที่ $\int_{|x-\mu|\geq\varepsilon}$ แทนอินทรีกรัลของตัวแปรสุ่ม X ในช่วง $|x - \mu| \geq \varepsilon$ และ $\int_{|x-\mu|<\varepsilon}$ แทนอินทรีกรัลของตัวแปรสุ่ม X ในช่วง $|x - \mu| < \varepsilon$ ดังไดอะแกรมต่อไปนี้



---

[1] การพิสูจน์จะดำเนินการกับเฉพาะกรณีที่ตัวแปรสุ่ม X เป็นตัวแปรสุ่มแบบต่อเนื่อง (Continuous Variate) เท่านั้น ถ้า X เป็นตัวสุ่มแบบตัดตอน (Discrete Variate) ก็สามารถพิสูจน์ได้และให้ผลลัพธ์ตรงกัน เพียงแต่ใช้เครื่องหมาย Σ แทนที่เครื่องหมาย $\int$ เท่านั้น

ดังนั้น

$$\sigma^2 \geq \int_{|x-\mu|\geq\varepsilon} (x-\mu)^2 f(x)dx$$: (หัก $\int_{|x-\mu|<\varepsilon}$ ออกได้เพราะอินทีแกรนด์เป็น Non-negative)

แต่ $(x - \mu)$ มีค่าแตกต่างกันได้อย่างน้อยเท่ากับ ε

ดังนั้น

$$\sigma^2 \geq \int_{|x-\mu|\geq\varepsilon} \varepsilon^2 f(x)dx$$

$$\Rightarrow \quad \int_{|x-\mu|\geq\varepsilon} f(x)dx \leq \frac{\sigma^2}{\varepsilon^2}$$

แต่อินทรีกัลของ $f(x)$ ในช่วง $|x - \mu| \geq \varepsilon$ หมายถึงความน่าจะเป็นที่ตัวแปรสุ่ม X จะอยู่ในช่วง $|x - \mu| \geq \varepsilon$

นั่นคือ

$$\int_{|x-\mu|\geq\varepsilon} f(x)dx = P\{|x-\mu| \geq \varepsilon\}$$

ดังนั้น

$$P\{|x-\mu| \geq \varepsilon \leq \frac{\sigma^2}{\varepsilon^2} \qquad \text{.....(ก)}$$

และเนื่องจาก $P\{|x - \mu| < \varepsilon\} = 1 - P\{|x - \mu| \geq \varepsilon\}$

ดังนั้น

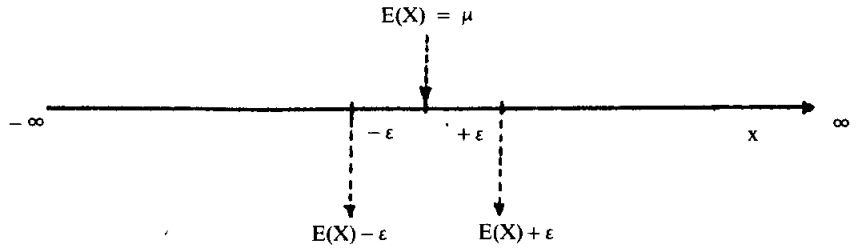
เมื่อนำผลในข้อ (ก) ไปหักออกจาก 1 ทั้งสองด้านของอสมการ (Inquality) จะได้

$$1 - P\{|x-\mu| \geq \varepsilon\} \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

$$\Rightarrow \quad P\{|x-\mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{\varepsilon^2} \qquad \text{.....(ข)}$$

พิสูจน์ วิธีที่ 2

พิจารณาไดอะแกรมต่อไปนี้



จะพบว่า

$$\sigma^2 = \int_{-\infty}^{\infty}(X - E(X))^2 f(x)dx$$

$$\geq \int_{-\infty}^{E(X)-\varepsilon}(X - E(X))^2 f(x)dx + \int_{E(X)+\varepsilon}^{\infty}(X - E(X))^2 f(x)dx$$

จาก $(X - E(X))$ กำหนดให้ X และ E(X) มีค่าต่างกันได้อย่างน้อยเท่ากับ ε

$$\Rightarrow \quad \sigma^2 \geq \int_{-\infty}^{E(X)-\varepsilon}\varepsilon^2 f(x)dx + \int_{E(X)+\varepsilon}^{\infty}\varepsilon^2 f(x)dx$$

$$\frac{\sigma^2}{\varepsilon^2} \geq \int_{-\infty}^{E(X)-\varepsilon} f(x)dx + \int_{E(X)+\varepsilon}^{\infty} f(x)dx$$

$$\geq \Pr\{X \leq E(X) - \varepsilon\} + \Pr\{X \geq E(X) + \varepsilon\}$$

$$\geq \Pr\{(X - E(X)) \leq -\varepsilon\} + \Pr\{(X - E(X)) \geq \varepsilon\}$$

$$\geq \Pr\{|X - E(X)| \geq \varepsilon\}$$

นั่นคือ

$$\Pr\{|x - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} \qquad \text{.....(ก)}$$

และในทำนองเดียวกันกับการพิสูจน์วิธีที่ 1 จะทำให้ได้ผลลัพธ์

$$\Pr\{|x - \mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{\varepsilon^2} \qquad \text{.....(ข)}$$

ตัวอย่าง 1.1 เราไม่อาจทราบได้ว่าความต้องการซื้อผลิตภัณฑ์ประเภทสบู่ของชุมชนแห่งหนึ่งมีการแจกแจงแบบใด ทราบแต่เพียงว่าในสัปดาห์หนึ่ง ๆ ชุมชนนั้นใช้สบู่เฉลี่ย 100 โหล ทั้งนี้มีค่าความแปรปรวนเท่ากับ 10 โหล จงคำนวณดูว่าในสัปดาห์ที่สองของเดือนชุมชนนั้นจะมีความต้องการซื้อสบู่ระหว่าง 90-110 โหล ด้วยความน่าจะเป็นเท่าไร

วิธีทำ ให้ตัวแปรสุ่ม X แทนความต้องการซื้อสบู่ของชุมชนแห่งนั้น

$$\Pr\{90 \leqslant X \leqslant 110\} = ?$$

พิจารณา $\Pr\{|X - E(X)| \leqslant \varepsilon\}$ จะพบว่า

$$\Pr\{|X - E(X)| \leqslant \varepsilon\} = \Pr\{-\varepsilon \leqslant X - E(X) \leqslant \varepsilon\}$$

$$= \Pr\{E(X) - \varepsilon \leqslant X \leqslant E(X) + \varepsilon\}$$

$$\Rightarrow \quad E(X) - \varepsilon = 90$$

และ $\quad E(X) + \varepsilon = 110$

$$\Rightarrow \quad E(X) = 100, \quad \varepsilon = 10$$

นั่นคือ

$$\Pr\{90 \leqslant X \leqslant 110\} = \Pr\{|X - 100| \leqslant 10\}$$

$$\geqslant 1 - \frac{\sigma^2}{\varepsilon^2}$$

$$\geqslant 1 - \frac{10}{10^2}$$

$$\geqslant 1 - .10$$

$$\geqslant .90$$

นั่นก็คือ โอกาสที่ชุมชนแห่งนั้นจะใช้ผลิตภัณฑ์ประเภทสบู่ในสัปดาห์ที่สองของเดือน (หรือสัปดาห์ใด ๆ) ระหว่าง 90 ถึง 100 โหล มีค่าไม่น้อยกว่า 90%

ตัวอย่าง 1.2 X และ Y เป็นตัวแปรสุ่มที่เป็นอิสระต่อกัน โดยที่ $V(X) = \frac{1}{4}$ และ $V(Y) = \frac{1}{2}$ จงคำนวณหา $\Pr\{|2(X - Y) - E(2(X - Y)| > 2.0\}$

วิธีทำ เนื่องจากตัวแปรสุ่ม X และ Y เป็นอิสระต่อกัน โดยที่ $V(X) = \frac{1}{4}$ และ $V(Y) = \frac{1}{2}$ และในการคำนวณ $\Pr\{|2(X - Y) - E\,2(X - Y)| > 2.0\}$ จะพบว่า ตัวแปรสุ่มที่มุ่งศึกษาในที่นี้คือ $2(X - Y)$ ซึ่งเป็นฟังก์ชั่นของ X และ Y จากอสมการของเชพบิเชฟ

$$\Pr\{|2(X - Y) - E\,2(X - Y)| > 2.0\} \leqslant \frac{V(2(X - Y))}{\varepsilon^2}$$

ในที่นี้ $\varepsilon = 2.0$ และ

$$V(2(X - Y)) = 2^2V(X - Y) = 4(V(X) + V(Y)) = 4(\frac{1}{4} + \frac{1}{2}) = 3;\ Cov(X, Y) = 0$$

ดังนั้น

$$\Pr\{|2(X - Y) - E\,2(X - Y)| > 2.0\} \leqslant \frac{3}{4}$$

$$\leqslant .75$$

ตัวอย่าง 1.3 โดยปกติเราทราบว่าจำนวนอุปกรณ์ที่เสื่อมคุณภาพในเครื่องรับโทรทัศน์เครื่องหนึ่ง ๆ ที่ใช้ไปแล้วระยะเวลาหนึ่ง (สมมุติว่าเป็นเวลา 3 ปี) มีการแจกแจงแบบพัวซอง (Poisson) โดยที่เมื่อใช้ถึงกำหนดเวลาดังกล่าวเครื่องรับโทรทัศน์เครื่องหนึ่งต้องเปลี่ยนอุปกรณ์โดยเฉลี่ยเครื่องละ 10 ชิ้น ถ้าสุ่มเครื่องรับโทรทัศน์ที่ใช้มาแล้วในช่วงเวลาดังกล่าวนั้นมา 1 เครื่อง

ก. จงหาความน่าจะเป็นที่เครื่องรับเครื่องนั้นจะต้องเปลี่ยนอุปกรณ์ประมาณ 6 ถึง 14 ชิ้น

ข. จงหาความน่าจะเป็นที่เครื่องรับเครื่องนั้นจะต้องเปลี่ยนอุปกรณ์ไม่น้อยกว่า 8 ชิ้น

วิธีทำ ให้ X เป็นจำนวนอุปกรณ์ของเครื่องรับโทรทัศน์ที่เสื่อมคุณภาพ

เราทราบว่าตัวแปรสุ่ม X มีการแจกแจงแบบพัวซอง ดังนั้น

$$E(X) = V(X) = \lambda = 10$$

ก. $\Pr\{6 \leqslant X \leqslant 14\} = ?$

$$\Pr\{6 \leqslant X \leqslant 14\} = \Pr\{6 - 10 \leqslant X - E(X) \leqslant 14 - 10\}$$

$$= \Pr\{-4 \leqslant X - E(X) \leqslant 4\}$$

$$= \Pr\{|X - E(X)| \leqslant 4\}$$

จะเห็นได้ว่า $\varepsilon = 4$ และ $V(X) = \lambda = 10$

ดังนั้น

$$\Pr\{6 \leqslant X \leqslant 14\} = \Pr\{|X - E(X)| \leqslant 4\}$$

$$\geqslant 1 - \frac{10}{16}$$

$$\geqslant .375$$

นั่นคือ โอกาสที่เครื่องรับโทรทัศน์ดังกล่าวจะต้องเปลี่ยนอุปกรณ์ระหว่าง 6 ถึง 14 ชิ้น มีได้ไม่น้อยกว่า .375

ข. Pr(เปลี่ยนอุปกรณ์ไม่น้อยกว่า 8 ชิ้น) = $Pr(X \geq 8)$ = ?

กรณีนี้ไม่อาจคำนวณหาค่าความน่าจะเป็นโดยอาศัยอสมการของเชพบิเชฟได้ เพราะไม่อาจจัด $Pr(X \geq 8)$ ให้เป็น Absolute Deviation ได้ ต้องคำนวณหาค่าโดยวิธีอื่น เช่นทฤษฎีการโน้มเข้าสู่เกณฑ์กลาง (Central Limit Theorem) ซึ่งจะกล่าวถึงในบทต่อไป

ตัวอย่าง 1.4 เครื่องอัดฝาจีบขวดน้ำอัดลมเครื่องหนึ่งสามารถผลิตฝาจีบได้ชั่วโมงละ 10,000 ฝา แต่ว่าประมาณ 1% ของฝาจีบที่ผลิตได้นั้นจะบิดเบี้ยวไม่อาจนำไปใช้งานได้ จงคำนวณหาค่า c ที่ในชั่วโมงแห่งการผลิตใด ๆ จำนวนฝาจีบที่ใช้การไม่ได้จะมีอยู่ทั้งสิ้นเพียง 100 − c ถึง 100 + c ฝาเท่านั้น ทั้งนี้ต้องเชื่อถือในความถูกต้องได้ถึง 95% เป็นอย่างน้อย

วิธีทำ ให้ p สัดส่วนของฝาจีบที่ใช้การไม่ได้ในชั่วโมงหนึ่ง ๆ

ดังนั้น $p = 1\% = .01$

$q = 1 - p = .99$ = สัดส่วนของฝาจีบที่มีสภาพสมบูรณ์

ในกรณีเช่นนี้ จะเห็นได้ว่าคุณภาพของฝาจีบแต่ละฝามีการแจกแจงแบบ Bernoulli และคุณภาพของฝาจีบที่ผลิตได้ในแต่ละชั่วโมงมีการแจกแจงแบบทวินาม

ให้ตัวแปรสุ่ม X แทนจำนวนฝาจีบที่ใช้การไม่ได้ในชั่วโมงแห่งการผลิตหนึ่ง

ดังนั้น เราจึงสามารถคาดได้ว่าในชั่วโมงหนึ่งจะผลิตฝาจีบที่ใช้การไม่ได้ $= np = (10,000)(.01) = 100$ ฝา ทั้งนี้ความแปรปรวน $= npq = (10,000)(.01)(.99) = 99$

สิ่งที่ต้องการคือ

$$Pr(100 - c < X < 100 + c) \geq .95, \quad c = ?$$

$$Pr(100 - c < X < 100 + c) \geq .95$$

$$Pr(-c < X - 100 < c) \geq .95 \quad ; \quad 100 = E(X) = np$$

$$Pr(|X - 100| \leq c) \geq .95 \qquad \text{.....(1)}$$

อาศัยอสมการของเชพบิเชฟ จะพบว่า $Pr(|X - 100| \leq c) \geq 1 - \frac{V(X)}{c^2} = 1 - \frac{99}{c^2}$

ดังนั้นสมการที่ (1) จะกลายเป็น

$$1 - \frac{99}{c^2} \geq .95$$

$$\frac{99}{c^2} \geq .05$$

$$\Rightarrow \quad c^2 \geq \frac{99}{.05}$$

$$\geq 1980$$

$$c \geq 45$$

นั่นคือ ในชั่วโมงแห่งการผลิตใด ๆ เครื่องจักรจะผลิตฝาจุกที่ใช้การไม่ได้ประมาณ 100 — 45 ถึง 100 + 45 55 ถึง 145 ฝา ทั้งนี้สามารถยืนยันความถูกต้องได้ถึง 95% เป็นอย่างน้อย

1.3 การกำหนดขนาดตัวอย่าง

ดังที่ได้กล่าวมาแล้วในตอนต้นว่า อสมการของเชพบิเชฟ นั้นนอกจากใช้คำนวณหาค่าขีดจำกัดบนและขีดจำกัดล่างของค่าความน่าจะเป็นที่ตัวแปรสุ่มใด ๆ จะมีค่าคลาดเคลื่อนไปจากค่าเฉลี่ยในขนาดของความแตกต่างหนึ่ง ๆ แล้ว เรายังสามารถใช้คำนวณหาขนาดของตัวอย่าง (Sample Size) เพื่อประโยชน์ในการสุ่มตัวอย่างอีกด้วย ต่อไปนี้จะแสดงตัวอย่างเพื่อให้เห็นประโยชน์ของอสมการของเชพบิเชฟในด้านนี้ แต่ก็ต้องขอให้ทำความเข้าใจไว้ในชั้นนี้ด้วยว่า ขนาดของตัวอย่างที่ได้โดยวิธีนี้ค่อนข้างหยาบ และดูจะเป็นขนาดตัวอย่างที่ใหญ่กว่าที่คำนวณได้จากวิธีอื่น ๆ ในประเด็นของปัญหาเดียวกัน ที่เป็นเช่นนี้ก็เพราะ อสมการของเชพบิเชฟให้ค่าต่าง ๆ ไม่ว่าค่าของความน่าจะเป็นหรือขนาดของตัวอย่าง โดยอาศัยข้อสนเทศจากตัวแปรเชิงสุ่มที่มุ่งศึกษานั้นน้อยมาก เมื่อได้รับข้อสนเทศน้อยคำตอบที่จะให้ได้ก็ย่อมออกมาในลักษณะ "เหวี่ยงแห" เป็นธรรมดา อุปมาดังการทำนายโชคชะตาราศรี ถ้าลูกค้าบอกแต่วันเกิด หมอดูก็ทายกันอย่างกว้าง ๆ ทายชนิด "เหวี่ยงแห" มันก็ต้องถูกทุกครั้งไป ถ้าได้ครบทั้งวันเดือนปีเกิดเวลาตกฟากตอนแพ้ท้องแม่อยากกินอะไร อย่างนี้หมอดูก็ทายกระชับ รัดกุม ไม่เหวี่ยงแหเหมือนก่อน ฉันใดก็ตาม การดำเนินทางสถิติถ้าได้ข้อสนเทศจากตัวแปรสุ่มที่มุ่งศึกษามากเพียงใด คำตอบที่จะได้ก็กระชับรัดกุมมากขึ้นเพียงนั้น แต่ต้องขอย้ำว่าอสมการของเชพบิเชฟนั้นมิใช่ว่าใช้ไม่ได้ผลความจริงใช้ได้ดีแต่จะดีเฉพาะเมื่อมีข้อสนเทศของตัวแปรที่มุ่งสนใจค่อนข้างน้อย เมื่อมีข้อสนเทศของตัวแปรที่มุ่งสนใจมากขึ้นก็ควรใช้วิธีการอื่น ๆ ซึ่งเหมาะสมกว่าสืบไป

ตัวอย่าง 1.5 บริษัทตัวแทนจำหน่ายนมสดกระป๋องแห่งหนึ่งสังเกตว่าร้านค้าย่อยที่รับนมสดไปจำหน่าย ส่งนมสดบูดรับประทานไม่ได้คืนมาบ่อยครั้ง เขามีความสงสัยและอยากทราบว่านมสดที่ผลิตจากบริษัทแล้ว ส่งมาให้นั้นมีกี่เปอร์เซนต์ที่เป็นนมบูด แต่ก็ไม่ทราบว่าจะสุ่มตัวอย่างนมมาสักกี่กระป๋องเพื่อตรวจสอบคุณภาพและไม่ทราบข้อเท็จจริงใด ๆ เกี่ยวกับเรื่องนี้เลย ต้องการทราบเพียงแต่ว่า จะสุ่มมากี่กระป๋องจึงจะเชื่อถือได้ถึง 95% เป็นอย่างน้อยว่าสัดส่วนของนมกระป๋องที่บูดจากกลุ่มตัวอย่าง (Sample Proportion) จะคลาดเคลื่อนไปจากสัดส่วนจริง (Population Proportion) จากการผลิตไม่ถึง 0.1

วิธีทำ ให้ $\hat{p}$ สัดส่วนของนมบูดจากกลุ่มตัวอย่าง

$\hat{q}$ สัดส่วนของนมที่ไม่บูดจากกลุ่มตัวอย่าง

p สัดส่วนจริงของนมบูดจากโรงงาน

$$\varepsilon = 0.1$$

ให้ตัวแปรสุ่ม X แทนคุณภาพของนมกระป๋องแต่ละกระป๋องโดยที่ $X = \begin{cases} 1 & \text{ถ้าบูด} \\ 0 & \text{ถ้าไม่บูด} \end{cases}$

ดังนั้น

$$\Pr\{|\hat{p} - p| < 0.1\} \geq .95$$

แต่ $$\hat{p} = \frac{1}{n}\sum_{i}^{n} X_i \; ; \quad X_i = 0, 1$$

และเนื่องจาก X_i มีผลเป็น 2 ลักษณะคือบูดหรือไม่บูด แสดงว่า X_i มีการแจกแจงแบบ Bernoulli

โดยมีค่า $E(X_i) = E(X) = p$ และ $V(X_i) = V(X) = pq$

ดังนั้น $S = \sum_{i}^{n} X_i$ จะมีการแจกแจงแบบ Binomial โดยมีค่า $E(S) = E(\sum_{i}^{n} X_i) = np$ และ $V(S) = V(\sum_{i}^{n} X_i) = npq$

ดังนั้น จากสมการ

$$\Pr\{|\hat{p} - p| < 0.1\} \geq .95$$

$$\Rightarrow \quad \Pr\{|\frac{1}{n}\sum_{i}^{n} X_i - p| < 0.1\} \geq .95$$

$$\Rightarrow \quad \Pr\{|\sum_{i}^{n} X_i - np| < 0.1n\} \geq .95 \qquad \text{.....(1)}$$

โดยที่ $np = E(\sum_{i}^{n} X_i)$

อาศัยอสมการของเชพบิเชฟ จะพบว่า

$$\Pr\{|\sum_{i}^{n} X_i - np| \leq 0.1\} \geq 1 - \frac{V(\sum_{i}^{n} X_i)}{(0.1n)^2} = 1 - \frac{npq}{(0.1n)^2}$$

แทนค่าลงในสมการที่ (1)

$$\Rightarrow \quad 1 - \frac{npq}{(0.1n)^2} \geq .95$$

$$1 - \frac{pq}{(0.1)^2 n} \geq .95$$

$$\frac{pq}{0.01n} \leq .05$$

$$n \geq \frac{pq}{(0.01)(.05)}$$

แต่ผลคูณ pq จะมีค่าสูงที่สุดเมื่อ $p = q = \frac{1}{2}$ และผลคูณจะมีค่าลดลงเมื่อ $p > \frac{1}{2}$ และ $p < \frac{1}{2}$

ดังนั้นขนาดตัวอย่างที่เหมาะสมคือขนาดตัวอย่างที่ได้จาก (ในกรณีนี้) สถานการณ์เมื่อผลคูณ pq มีค่าสูงสุด นั่นคือ

$$n \geq \frac{(\frac{1}{2})(\frac{1}{2})}{(.01)(.05)} \quad \Rightarrow \quad n \geq 500$$

นั่นคือ ควรสุ่มตัวอย่างนมสดมาตรวจสอบคุณภาพอย่างน้อย 500 กระป๋อง จึงจะสามารถคาดหมายได้ว่าสัดส่วนที่ได้จากกลุ่มตัวอย่างคลาดไปจากสัดส่วนจริงไม่ถึง 10%

ตัวอย่าง 1.6 กรมสรรพาวุธทหารกำลังทดลองประสิทธิภาพของขีปนาวุธชนิดใหม่ จากการศึกษาพบว่าขีปนาวุธจากต้นแบบเดียวกันซึ่งผลิตในประเทศเยอรมันสามารถยิงได้ไกลถึง 6,000 หลาโดยเฉลี่ย มีความเบี่ยงเบนมาตรฐาน 10 หลา อยากทราบว่าขีปนาวุธประเภทเดียวกันซึ่งผลิตโดยกรมสรรพาวุธทหารนี้จะต้องทำการปรับปรุงถึงกี่ครั้ง (หมายความว่า เมื่อทดลอง 1 ครั้ง ระยะยิงไม่ถึงมาตรฐานของต้นแบบ ก็ต้องนำไปปรับปรุงกันครั้งหนึ่ง) จึงจะเชื่อถือได้ถึง 99% เป็นอย่างน้อยว่า ขีปนาวุธของกรมสรรพาวุธมีระยะยิงเฉลี่ยคลาดเคลื่อนไปจากระยะมาตรฐานของต้นแบบไม่ถึง 20 หลา

วิธีทำ

สิ่งที่ต้องการคือ จากสมการ

$$\Pr\{|\bar{X} - E(\bar{X})| < 20\} \geq .99 \quad ; \quad n = ?$$

โดยอาศัย อสมการของเชพบิเชฟ จะพบว่า

$$\Pr\{|\bar{X} - E(\bar{X})| < 20\} \geq 1 - \frac{V(\bar{X})}{(20)^2}$$

$$\geq 1 - \frac{\sigma^2}{n(20)^2}$$

$$1 - \frac{\sigma^2}{n(20)^2} \geq .99$$

$$\Rightarrow \quad \frac{\sigma^2}{n(20)^2} \leq .01$$

$$n \geq \frac{\sigma^2}{(.01)(400)}$$

$$\geq \frac{100}{(.01)(400)} \quad ; \quad \sigma^2 = 100$$

$$\geq 25$$

แสดงว่า ถ้ากรมสรรพาวุธจะสร้างขีปนาวุธโดยที่แน่ใจได้อย่างน้อย 99% ว่าขีปนาวุธดังกล่าวจะมีคุณภาพใกล้เคียงกับต้นแบบ มีระยะยิงพลาดไปจากต้นแบบไม่ถึง 20 หลา กรมสรรพาวุธจะต้องปรับปรุงทางเทคนิคถึง 25 ครั้งเป็นอย่างน้อย

1.4 กฎการใช้กลุ่มตัวอย่างขนาดใหญ่ (Law of Large Number)

ให้ $X_1, X_2, \ldots$ เป็นชุดของตัวแปรสุ่มใด ๆ ซึ่งมีค่าคาดหมายเป็น $E(X_1), E(X_2), \ldots$ ตามลำดับ สมมุติตัวแปรสุ่ม $Y = \sum_{i=1}^{n} X_i$ มีค่าความแปรปรวนปรากฏ (exist) ดังนั้น เราจะได้กฎของการใช้กลุ่มตัวอย่างขนาดใหญ่ดังนี้

ทฤษฎี 1.2 ถ้า $V(\frac{1}{n} \sum_{i}^{n} X_i)$ มีค่าเข้าใกล้ 0 ขณะที่ n มีขนาดใหญ่ขึ้น $(n \to \infty)$ และถ้าให้ ε เป็นจำนวนเต็มบวกใด ๆ

แล้ว

$$\Pr\{|\frac{1}{n} \sum_{i}^{n}(X_i - E(X_i)| \geq \varepsilon\} \to 0 \quad \text{เมื่อ } n \to \infty \qquad \ldots\ldots\ldots(1)$$

หรือกล่าวได้ว่า

ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่แล้วความคลาดเคลื่อนโดยเฉลี่ยระหว่างค่าสังเกต (X_i) กับค่าคาดหมาย $(E(X_i))$ จะมีโอกาสเกิดขึ้นได้น้อยมาก

หรืออีกนัยหนึ่ง

$$\Pr\{|\frac{1}{n} \sum_{i}^{n}(X_i - E(X_i))| < \varepsilon\} \to 1 \quad \text{เมื่อ } n \to \infty \qquad \ldots\ldots\ldots(2)$$

ซึ่งกล่าวเป็นคำพูดได้ว่า

ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่แล้ว ความคลาดคลึงกันโดยเฉลี่ยระหว่างค่าสังเกตและค่าคาดหมายจะมีโอกาสเกิดขึ้นได้มาก

หมายเหตุ การตีความกฎการใช้กลุ่มตัวอย่างขนาดใหญ่ ความหมายที่ (1) และ (2) อาจจะยังคลุมเคลืออยู่ จะลองแปลงรูปทั้ง (1) และ (2) แล้วตีความให้ดูใหม่ ซึ่งอาจทำให้เข้าใจได้กระจ่างขึ้น

จาก (1)

$$\Pr\{|\frac{1}{n} \sum_{i}^{n}(X_i - E(X_i)| \geq \varepsilon\} \to 0 \qquad \text{เมื่อ } n \to \infty$$

แปลงรูปเป็น

$$\Pr\{|\frac{1}{n} \sum_{i}^{n} X_i - \frac{1}{n} n.E(X)| \geq \varepsilon\} \to 0 \qquad \text{เมื่อ } n \to \infty ; \because E(X_i) = E(X) = \mu$$

$$\Pr\{|\bar{x} - \mu| \geq \varepsilon\} \to 0 \text{ เมื่อ } n \to \infty$$

ซึ่งก็ตีความได้ว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ ($n \to \infty$) แล้วโอกาสที่ค่าเฉลี่ยจากกลุ่มตัวอย่างจะคลาดไปจากค่าเฉลี่ยจริงแม้แต่เพียงเล็กน้อยแทบจะไม่มีเลย หรือนัยหนึ่งค่าเฉลี่ยจากตัวอย่างสามารถนำไปใช้ประมาณค่าเฉลี่ยจริงได้อย่างแม่นยำ ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่พอ

และจาก (2) จะได้ว่า

$$\Pr\{|\bar{x} - \mu| < \varepsilon\} \to 1 \quad \text{เมื่อ} \quad n \to \infty$$

หมายความว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่แล้ว โอกาสที่ค่าเฉลี่ยจากกลุ่มตัวอย่างและค่าเฉลี่ยจริงจะมีค่าเท่ากันได้นั้นจะมีได้อย่างสมบูรณ์เกือบ 100%

พิสูจน์ ให้ตัวแปรสุ่ม $W = \frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))$

$$E(W) = E(\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))) = \frac{1}{n}\sum_{i}^{n}E(X_i - E(X_i)) = 0$$

ก. อาศัยอสมการของเชพบิเชฟ พบว่า

$$\Pr\{|W - E(W)| \geq \varepsilon\} \leq \frac{V(W)}{\varepsilon^2}$$

$$\Rightarrow \quad \Pr\{|\frac{1}{n}\sum_{i}^{n}(X_i - E(X)) - 0| \geq \varepsilon\} \leq \frac{V(\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))}{\varepsilon^2}$$

$$\Rightarrow \quad \Pr\{|\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i)| \geq \varepsilon\} \leq \frac{V(\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))}{\varepsilon^2}$$

$$\because \quad V(\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i)) = \frac{1}{n^2}\sum_{i}^{n}V(X_i - E(X_i))$$

$$= \frac{1}{n^2}\sum_{i}^{n}V(X_i) - \frac{1}{n^2}\sum_{i}^{n}V(E(X_i))$$

$$= \frac{1}{n^2}\sum_{i}^{n}V(X_i) - 0 \quad \because E(X_i) = \mu_i \text{ เป็นตัวคงที่ และ Variance ของตัวคงที่มีค่าเป็น 0}$$

$$\Rightarrow \quad \Pr\{|\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))| \geq \varepsilon\} \leq \frac{\sum_{i}^{n}V(X_i)}{n^2\varepsilon^2}$$

$$\therefore \quad \lim_{n\to\infty} \Pr\{|\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))| \geq \varepsilon\} \leq \lim_{n\to\infty} \frac{\sum_{i}^{n} V(X_i)}{n^2\varepsilon^2} \to 0$$

นั่นคือ $\Pr\{|\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))| \geq \varepsilon\} \to 0$ เมื่อ $n \to \infty$

ในทำนองเดียวกัน เราสามารถพิสูจน์ได้ว่า

$$\Pr\{|\frac{1}{n}\sum_{i}^{n}(X_i - E(X_i))| < \varepsilon\} \to 1 \quad \text{เมื่อ } n \to \infty$$

บทแทรก 1.1 ให้ตัวแปรสุ่ม $X_1, X_2, ..., X_n$ เป็นอิสระต่อกัน มีค่าคาดหมายและความแปรปรวนร่วมกันคือ[1] คือ $E(X_1) = E(X_2) = ... = E(X_n) = \mu$ และ $V(X_1) = V(X_2) = ... = V(X_n) = \sigma^2$ ดังนั้น ถ้า ε เป็นจำนวนเต็มบวกใด ๆ $(\varepsilon > 0)$ จะได้

$$\lim_{n\to\infty} \Pr\{|\bar{x} - \mu| \geq \varepsilon\} = 0$$

พิสูจน์ จาก $\Pr(|\bar{x} - \mu| \geq \varepsilon\}$

อาศัยอสมการของเชพบิเชฟ

$$\Rightarrow \quad \Pr\{|\bar{x} - \mu| \geq \varepsilon\} \leq \frac{V(\bar{X})}{\varepsilon^2}$$

$$\leq \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow \quad \lim_{n\to\infty} \Pr\{|\bar{x} - \mu| \geq \varepsilon\} \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} \to 0$$

นั่นคือ $\Pr\{|\bar{x} - \mu| \geq \varepsilon\} \to 0$ เมื่อ $n \to \infty$

หมายความว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ในการสุ่มตัวอย่างจากกลุ่มประชากร ค่าเฉลี่ยจากกลุ่มตัวอย่าง $(\bar{x})$ และค่าเฉลี่ยจริง ๆ (μ) จะมีโอกาสคลาดกันน้อยมาก

---

[1] หมายความว่า Sampled Random Variable $X_1, X_2, ..., X_n$ จากกลุ่มประชากร X ซึ่งมีค่าคาดหมายและความแปรปรวนเป็น μ และ σ^2 ตามลำดับ

บทแทรก 1.2 Bernoulli's Thorem ถ้า S_n เป็นยอดรวมของ Success จากการทดลองที่เป็นอิสระต่อกัน n ครั้ง ที่แต่ละครั้งของการทดลองมีโอกาสประสบความสำเร็จเท่ากับ p และถ้า ε เป็นจำนวนเต็มบวกใด ๆ ($\varepsilon > 0$) แล้ว

$$\Pr\{ \frac{1}{n} S_n - P| \geq \varepsilon\} \rightarrow 0 \quad \text{เมื่อ } n \rightarrow \infty$$

และ

$$\Pr\{| \frac{1}{n} S_n - p| < \varepsilon\} \rightarrow 1 \quad \text{เมื่อ } n \rightarrow \infty$$

หมายเหตุ กรณีของ Bernoulli Trial นั้น ตัวแปรสุ่มแต่ละตัวจะมีค่าได้เพียง 2 ค่าอย่างไรอย่างหนึ่งเท่านั้น คือ 0 และ 1 อาจเป็น 0 ก็ต่อเมื่อ Success และ 1 เมื่อ Fail หรือกลับกัน ดังนั้น

$\frac{S_n}{n} = \frac{\sum_i^n X_i}{n} = \hat{p}$ = สัดส่วนหรือเปอร์เซนต์ของ Success (หรือ Fail ก็ได้แล้ว แต่ว่าเราจะกำหนดให้ 0 และ 1 หมายถึงอะไร) จะเห็นได้ว่า $\hat{p}$ เป็นกรณีของเฉพาะ $\bar{x}$

พิสูจน์

จาก $\Pr\{| \frac{1}{n} S_n - p| \geq \varepsilon\}$

$=>$ $\Pr\{|S_n - np| \geq n\varepsilon\}$

แต่ $S_n = \sum_i^n X_i$ และ X_i มีการแจกแจงแบบ Bernoulli ดังนั้น $S_n = \sum_i^n X_i$ จะมีการแจกแจงแบบ Binomial มีค่าเฉลี่ยเป็น $E(S_n) = np$ และมีความแปรปรวน $V(S_n) = npq$ ดังนั้น โดยอาศัยอสมการของเชพบิเชฟ

$$=> \quad \Pr\{|S_n - np| \geq n\varepsilon\} \leq \frac{V(S_n)}{(n\varepsilon)^2}$$

$$\leq \frac{npq}{(n\varepsilon)^2}$$

$$\leq \frac{pq}{n\varepsilon^2}$$

$$=> \quad \lim_{n\rightarrow\infty} \Pr\{|S_n - np| \geq \varepsilon\} \leq \lim_{n\rightarrow\infty} \frac{pq}{n} \rightarrow 0$$

นั่นคือ $\Pr\{|\frac{1}{n} S_n - p| \geq \varepsilon\} \to 0$ เมื่อ $n \to \infty$ หรือกล่าวได้ว่า เมื่อดำเนินการสุ่มตัวอย่างขนาดใหญ่ จะพบว่า $\hat{p}$ จะมีโอกาสคลาดไปจาก p ได้น้อยมาก หรือ นัยหนึ่งเราสามารถนำค่าสัดส่วนหรือเปอร์เซนต์ที่ได้จากกลุ่มตัวอย่างไปใช้ประมาณค่าสัดส่วนหรือเปอร์เซนต์จริงได้อย่างแม่นยำ ถ้าการสุ่มตัวอย่างนั้นใช้กลุ่มตัวอย่างขนาดใหญ่ ในทำนองเดียวกัน เราสามารถอาศัยอสมการของเชพบิเชฟ พิสูจน์ได้อีกว่า

$$\Pr\{|\frac{1}{n} S_n - p| < \varepsilon\} \to 1 \quad \text{เมื่อ } n \to \infty$$

ขอให้สังเกตไว้ ณ ที่นี้ว่า โดยข้อเท็จจริงแล้วกฎการใช้กลุ่มตัวอย่างขนาดใหญ่ เป็นเพียงส่วนหนึ่งหรือประโยชน์เพียงส่วนหนึ่งของอสมการของเชพบิเชฟเท่านั้น การนำไปใช้ประโยชน์ เช่น การหาขนาดตัวอย่างที่เหมาะสมกับเงื่อนไขที่เราต้องการ จึงคำนวณได้ในลักษณะเดียวกัน อาจมีบางคนแย้งว่า เมื่อการนำไปใช้ประโยชน์มิได้ต่างไปจาก อสมการของเชพบิเชฟ หรือเอาอสมการของเชพบิเชฟไปใช้โดยตรง แล้วจะศึกษาเรื่องกฎการใช้กลุ่มตัวอย่างขนาดใหญ่ด้วยวัตถุประสงค์ใดกัน ข้อนี้ขอชี้แจงว่า กฎการใช้กลุ่มตัวอย่างขนาดใหญ่นั้นเป็นเพียงกฎเกณฑ์ที่บอกให้เราทราบว่า "เมื่อจะทำการวิจัยหรือเก็บขนาดตัวอย่างเพื่อนำมาวิจัย ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่ หรือเก็บตัวอย่างมาเป็นจำนวนมากเพียงใด ค่าที่ได้จากกลุ่มตัวอย่าง (Statistics) จะมีค่าใกล้เคียงกับค่าจริง (ซึ่งโดยปกติเป็นค่าที่เราไม่อาจทราบได้) มากเพียงนั้น" ยกตัวอย่าง เช่น เราต้องการทราบว่าหลอดสูญญากาศที่ใช้ในอุปกรณ์ไฟฟ้า ที่ผลิตจากโรงงานแห่งหนึ่งจะมีอายุใช้งานนานเพียงใด ต่อปัญหานี้ อายุการใช้งานของหลอดสูญญากาศคือค่าจริง (Parameter) หรืออายุจริงซึ่งทุกคนอยากทราบ แต่ก็ไม่อาจทราบได้การจะทราบก็ต้องทำการทดลองสุ่มตัวอย่างหลอดสูญญากาศมาจำนวนหนึ่งแล้วทดสอบการทำงาน อาจดำเนินการโดยผ่านกระแสไฟฟ้าเข้าไปตลอดเวลา แล้วบันทึกเวลาไว้ว่าตั้งแต่เริ่มผ่านกระแสไฟฟ้าเข้าไปจนกระทั่งหลอดขาดหรือเสื่อมคุณภาพนั้นใช้เวลากี่ชั่วโมงหรือกี่วัน หรืออาจใช้เทคนิคอย่างอื่นตามความเหมาะสมทางเทคนิคก็ได้ การดำเนินการดังกล่าวนั้นถ้านำหลอดสูญญากาศมาทดสอบเป็นจำนวนน้อย ผลลัพธ์ก็ไม่ค่อยน่าเชื่อถือ แต่ถ้านำหลอดเป็นจำนวนมากมาทดสอบ ผลลัพธ์ (ซึ่งอาจหมายถึงอายุเฉลี่ย) จะน่าเชื่อถือกว่า และเป็นที่เชื่อถือได้มากกว่าผลลัพธ์จากการทดลองกับกลุ่มตัวอย่างขนาดใหญ่กว่านั้นมีค่าใกล้กับค่าอายุการใช้งานจริงของหลอดชนิดดังกล่าว ดังนี้เป็นต้น

ตัวอย่าง 1.7 อยากทราบว่าเราควรจะทอดลูกเต๋าสักกี่ครั้งจึงจะเป็นที่น่าเชื่อถือได้ถึง 95% เป็นอย่างน้อยว่า สัดส่วน (Proportion) ของการเกิดหน้าเอี่ยว (หน้าหนึ่ง) จะคลาดไปจากสัดส่วนจริงอย่างมากที่สุดเพียง 0.01

วิธีทำ

ค่าสัดส่วนจากกลุ่มตัวอย่าง $= \frac{1}{n} S_n$

ค่าสัดส่วนจริง $= p = \frac{1}{6} = .1667$

สิ่งที่ต้องการคือ

$$\Pr\{|\frac{1}{n} S_n - p| \leqslant 0.01\} \geqslant .95 \qquad n = ?$$

จะพบว่า

$$n \geqslant \frac{pq}{(.01)^2(.05)} \quad ; \quad q = 1 - p$$

$$n \geqslant 27{,}776$$

นั่นคือ ถ้าทำการทอดลูกเต๋าตั้งแต่ 27,776 ครั้งขึ้นไปแล้วเราจะเชื่อได้ถึง 95% เป็นอย่างน้อยว่าสัดส่วนของการหงายหน้าเอี่ยวที่ได้จากการทดลองจะคลาดเคลื่อนไปจากค่าสัดส่วนจริงไม่เกิน 0.01

สิ่งหนึ่งที่น่าสังเกตก็คือ ในอสมการ $n \geqslant \frac{pq}{(.01)^2(.05)}$ หรือเขียนเป็นรูปทั่ว ๆ ไปได้เป็น $n \geqslant \frac{pq}{\varepsilon^2\delta}$ นั้น ถ้าเป็นสถานการณ์ที่เป็นการนำไปประยุกต์ในงานต่าง ๆ เช่น การสำรวจประชามติ การตรวจสอบคุณภาพของผลิตภัณฑ์ทางอุตสาหกรรมหรือผลผลิตทางการเกษตร การวิจัยทางสังคม การวิจัยตลาด หรืออื่น ๆ ตัว p จะเป็นตัวไม่ทราบค่า (Unknown Paramete) จึงเป็นที่น่าสงสัยว่าจะคำนวณหาขนาดตัวอย่าง n ได้อย่างไร ในสถานการณ์เช่นนั้น

ในกรณีเช่นนี้ เราสามารถนำเอาข้อเท็จจริงที่เกี่ยวกับค่าสูงสุดของผลคูณ pq มาใช้ได้ กล่าวคือ ถ้าจะให้กล่าวได้อย่างรัดกุมและปลอดภัยแล้ว ควรใช้ค่าของ p และ q ที่ทำให้ผลคูณ pq มีค่าสูงสุด ค่าผลคูณดังกล่าวจะทำให้ค่าของขนาดตัวอย่าง n ที่ได้มีขนาดใหญ่ที่สุดในบรรดาค่าที่เป็นไปได้ทุกค่าของ p ขอให้พิจารณาตารางเปรียบเทียบต่อไปนี้ ซึ่งอาจแสดงให้เห็นถึงขนาดตัวอย่างที่แปรไปตามค่าของ p และเพื่อความสะดวกจะใช้ $\varepsilon = .01$ และ $\delta = .05$ เช่นเดิมดังตัวอย่าง 1.7

p	.1	.2	.3	.4	.5	.6	.7	.8	.9
q	.9	.8	.7	.6	.5	.4	.3	.2	.1
pq	.09	.16	.21	.24	.25	.24	.21	.16	.36
n	18,000	32,000	42,000	48,000	50,000	48,000	42,000	32,000	18,000

หมายเหตุ $n \geq \frac{pq}{(.01)^2(.05)}$

จะเห็นได้ว่า เมื่อ $p = q = \frac{1}{2}$ จะให้ค่าผลคูณ pq มีค่าสูงที่สุด และขนาดตัวอย่าง n ที่ได้จากค่านี้ จะเป็นขนาดตัวอย่างที่ใหญ่ที่สุดในระหว่างค่า p ที่เป็นไปได้ทั้งหมด ซึ่งเมื่อพิจารณากันในแง่ของทฤษฎีการใช้กลุ่มตัวอย่างขนาดใหญ่ ก็จะเห็นได้ทันทีว่ากลุ่มตัวอย่างที่เลือกขึ้นมาใช้ในกรณีที่ไม่ทราบค่าของ p นั้น ถ้าใช้ในกรณี $p = q = \frac{1}{2}$ ค่าที่พึงประเมินจะได้รับจากกลุ่มตัวอย่างจะเป็นค่าที่ใกล้เคียงกับค่าจริงมากที่สุด (พลิกไปดูตัวอย่าง 1.5)

ตัวอย่าง 1.8 ในการวิจัยเพื่อศึกษาอัตราการเกิดอุบัติเหตุบริเวณสี่แยกทุกแห่งในเขตกรุงเทพมหานครระหว่างเวลา 08.00 ถึง 10.00 น. ผู้วิจัยทราบแต่เพียงว่าสัมประสิทธิ์ของความผันแปร (C.V) มีค่าเท่ากับ .10 (จากสถิติกรมตำรวจ) แต่ไม่ทราบว่าจะดำเนินการสำรวจกี่ครั้ง (วัน) จึงจะเชื่อถือได้ถึง 90% เป็นอย่างน้อยว่า อัตราเฉลี่ยของการเกิดอุบัติเหตุในช่วงเวลาดังกล่าวจะมีค่าใกล้เคียงกับอัตราที่พึงปรากฏจริงมากที่สุด โดยกำหนดไว้ว่า ค่าที่ได้รับจากผลการวิจัยจะต้องคลาดไปจากอัตราที่ปรากฏจริงไม่เกิน 5% ถ้าท่านเป็นนักสถิติ ท่านจะให้ความช่วยเหลือนักวิจัยผู้นั้นได้อย่างไร

วิธีทำ เนื่องจากการเกิดอุบัติเหตุหรือปรากฏการณ์ใด ๆ ภายในช่วงเวลาที่กำหนดมีการแจกแจงแบบ Poisson[1]

ให้ตัวแปรสุ่ม X แทนจำนวนอุบัติเหตุที่ปรากฏขึ้นระหว่างเวลา 8.00 น. ถึง 10.00 น. ในบริเวณสี่แยกทุกแห่งในกรุงเทพฯ

ดังนั้น $f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$; $x = 0, 1, 2, \ldots$

[1] จะกล่าวถึงสถานการณ์ของการแจกแจงของตัวแปรสุ่มโดยละเอียดในบทที่ 4

โดยมีค่าคาดหมาย และความแปรปรวนเป็น

$E(X) = \lambda$ และ $\sigma^2 = V(X) = \lambda$ ตามลำดับ

เมื่อ $E(X) = \lambda =$ อัตราเฉลี่ยของการเกิดอุบัติเหตุต่อวันในระหว่างเวลา 08.00 ถึง 10.00 น.

ตามข้อกำหนด $CV = .10$ นั่นคือ $CV = \frac{\sigma}{\mu} = \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}}$

สิ่งที่ต้องการคือ จากอสมการ

$$\Pr\{|\bar{x} - \lambda| \leq .05\,\lambda\} \geq .90 \; ; \; n = ?$$

จาก $\Pr\{|\bar{x} - \lambda| \leq .05\,\lambda\}$ เมื่ออาศัย อสมการของเชพบิเชฟ จะพบว่า

$$\Pr\{|\bar{x} - \lambda| \leq .05\,\lambda\} \geq 1 - \frac{V(\bar{X})}{(.05\lambda)^2}$$

$$\geq 1 - \frac{\sigma^2}{n(.05\lambda)^2}$$

$$\geq 1 - \frac{\lambda}{n(.05)^2} \cdot \frac{1}{\lambda^2}$$

$$\geq 1 - \frac{1}{n(.05)^2} \cdot (CV)^2 = 1 - \frac{(.10)^2}{n(.05)^2}$$

$$\Rightarrow \quad 1 - \frac{(.10)^2}{n(.05)^2} \geq .90$$

$$n \geq \frac{(.10)}{(.05)^2(1.0)}$$

$$\geq 40$$

แสดงว่า นักวิจัยผู้นั้นจะต้องดำเนินการสำรวจเพื่อเก็บรวบรวมข้อมูลอย่างน้อย 40 ครั้ง (วัน) ในเวลา 08.00 น. ถึง 10.00 น. จึงจะเชื่อถือได้ถึง 90% เป็นอย่างน้อยว่า อัตราการเกิดอุบัติเหตุที่ได้จากการวิจัย (ผลการวิจัย) คลาดไปจากที่ปรากฏจริงอยู่สูงสุดเพียง 5% เท่านั้น