ความน่าจะเป็น กับการประยุกต์ใช้ ในการอนุมานเชิงสถิติ

Probability with Application to Statistical Inference

ฉบับปรับปรุง

มานัดดุ์ คำกอง

ภาควิชาสถิติ คณะวิทยาศาสตร์

มหาวิทยาลัยเชียงใหม่

สารบัญ

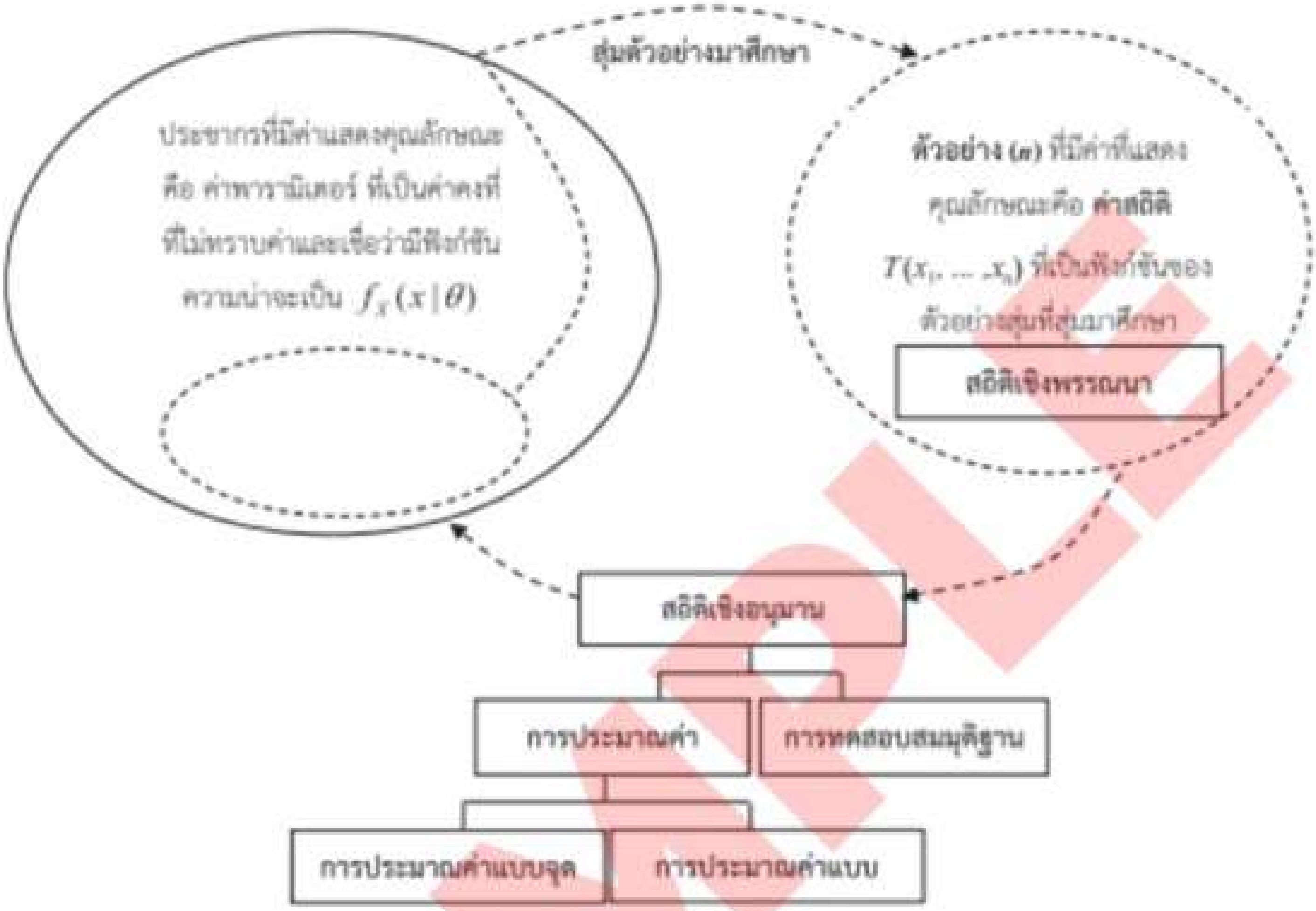
คำนำ (ฉบับปรับปรุง)	III
คำนำ	V
สารบัญ	VII
สารบัญตาราง	XII
สารบัญภาพ	XIII

บทที่ 1 ความน่าจะเป็นและสถิติ 1

1.1	บทนำ	2
1.2	ความน่าจะเป็นและเทคนิคการนับ	9
1.3	แนวคิดพื้นฐานความน่าจะเป็นและสมบัติความน่าจะเป็น	18
	แบบฝึกหัด	29
	บรรณานุกรม	32

บทที่ 2 ตัวแปรสุ่ม ฟังก์ชันความน่าจะเป็นและสมบัติที่สำคัญ 33

2.1	บทนำ	34
2.2	ตัวแปรสุ่มและฟังก์ชันความน่าจะเป็น	35
2.3	การคาดหมายและความแปรปรวน	44
2.4	โมเมนต์และฟังก์ชันก่อกำเนิดโมเมนต์	55
	แบบฝึกหัด	65
	บรรณานุกรม	72



ภาพที่ 1.1 ขอบข่ายของการศึกษาสถิติศาสตร์

ก่อนที่จะศึกษาขอบข่ายของศาสตร์สถิติจะอธิบายคำนิยามศัพท์ที่เกี่ยวข้องกับสถิติ ดังนี้

1) ประชากร (population) หมายถึง ชุดขององค์ประกอบหรือขอบเขตทั้งหมดที่สนใจศึกษา อาจเป็นคน สัตว์ หรือสิ่งของก็ได้ สามารถแบ่งเป็น 2 ประเภท คือ

1.1) ประชากรที่ประกอบด้วยชุดขององค์ประกอบหรือขอบเขตทั้งหมดที่สนใจศึกษามีจำนวนนับที่แน่นอน เรียกว่า ประชากรอันตะ (finite population) เช่น จำนวนพนักงานที่ปฏิบัติงานในโรงงานผลิตชิ้นส่วนประกอบยานยนต์ชนิดหนึ่งในช่วงเดือนธันวาคม พ.ศ. 2559 จำนวนคำที่ปรากฏในหนังสือความน่าจะเป็นกับการประยุกต์ใช้ในการอนุมานเชิงสถิติ เป็นต้น

1.2) ประชากรที่ประกอบด้วยชุดขององค์ประกอบหรือขอบเขตทั้งหมดที่สนใจศึกษามีจำนวนนับที่ไม่แน่นอน เรียกว่า ประชากรอนันต์ (infinite population) เช่น จำนวนปลาที่อาศัยอยู่ในทะเลทั่วโลก จำนวนนกที่อาศัยอยู่ในประเทศไทย เป็นต้น

10) สถิติเชิงอนุมาน (inferential statistics) หมายถึง การใช้ระเบียบวิธีการทางสถิติเพื่อจัดทำข้อสรุปของตัวอย่างที่สุ่มมาศึกษาเพื่ออนุมานไปสู่ประชากรโดยการประเมินความน่าเชื่อถือของวิธีการของสถิติดังกล่าวนั้น ๆ

ตัวอย่างที่ 1.1 นักวิจัยสิ่งแวดล้อมท่านหนึ่งได้รับข้อร้องเรียนจากชาวบ้านเกี่ยวกับโรงงานผลิตสารกำจัดศัตรูพืชแห่งหนึ่งที่ตั้งอยู่ใกล้กับแม่น้ำได้ลักลอบปล่อยน้ำเสียลงสู่แม่น้ำ ทำให้สิ่งมีชีวิต ที่อาศัยอยู่ในแม่น้ำตายเป็นจำนวนมากและแม่น้ำเน่าเสียไม่สามารถอุปโภคได้ นักวิจัยท่านนี้ จึงได้สุ่มตรวจระบบบำบัดน้ำเสียของโรงงานแห่งนี้ปริมาณ 500 มิลลิลิตร เพื่อตรวจวัดว่าค่า pH, BOD, SS, FOG และ TDS จากน้ำเสียที่ปล่อยจากโรงงานแห่งนี้มาจำนวน 30 วัน จะมีค่าเกินกว่าที่คณะกรรมการควบคุมมลพิษกำหนดหรือไม่ ดังนี้

ค่าวัด	ค่ามาตรฐาน
• pH คือ ค่าความเป็นกรด-ด่าง	• อยู่ระหว่าง 5.5 – 9 pH meter
• BOD (Biochemical Oxygen Demand) คือ ปริมาณของออกซิเจนที่แบคทีเรียใช้ย่อยสลายสารอินทรีย์ใน 5 วัน ที่ 20 องศาเซลเซียส	• ไม่เกิน 20 มก/ล หรือขึ้นอยู่กับประเภทอุตสาหกรรมแต่ไม่เกิน 60 มก/ล
• SS (Suspended Solids) คือค่าสารแขวนลอย	• ไม่เกิน 50 มก/ล
• FOG (Fat, oil and Grease) คือ ค่าน้ำมันและไขมัน	• ไม่เกิน 5 มก/ล หรือไม่เกิน 15 มก/ล
• TDS (Total Dissolved Solids) คือ ของแข็งที่ละลายได้ทั้งหมด	• ไม่เกิน 3,000 มก/ล หรือไม่เกิน 5,000 มก/ล

ที่มา: กรมควบคุมมลพิษ กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม

หน่วยวัด: มิลลิกรัมต่อลิตร (มก/ล)

ประชากรที่สนใจศึกษา คือ ปริมาณน้ำเสียที่ปล่อยจากโรงงานแหล่งนี้ทั้งหมด

ตัวอย่าง คือ ปริมาณน้ำเสียที่ปล่อยจากโรงงานแหล่งนี้ที่สุ่มมาตรวจในแต่ละวัน ปริมาณ 500 มิลลิลิตร จำนวน 30 วัน

หน่วยตัวอย่าง คือ ปริมาณน้ำเสียที่ปล่อยจากโรงงานแหล่งนี้ในแต่ละวัน

หน่วยการแจงนับ คือ เนื่องจากปริมาณน้ำเสียที่ปล่อยจากโรงงานแหล่งนี้สามารถวัดค่าต่าง ๆ ที่สนใจศึกษาได้ เช่น pH, BOD, SS, FOG และ TDS ดังนั้น หน่วยการแจงนับจะเหมือนกับหน่วยตัวอย่าง

สำหรับการคำนวณความน่าจะเป็นของเหตุการณ์ สามารถจำแนกได้เป็น 3 แนวทาง ดังภาพที่ 1.2

ภาพที่ 1.2 ชนิดของความน่าจะเป็น

1) ความน่าจะเป็นเชิงคลาสสิก (classical probability) หรือบางทีเรียกว่าความน่าจะเป็นก่อน (prior probability) คือ ความน่าจะเป็นของเหตุการณ์ที่สนใจศึกษาในปริภูมิตัวอย่าง Ω ของการทดลองสุ่ม โดยที่โอกาสในการเกิดเหตุการณ์ต่าง ๆ จะมีโอกาสเกิดขึ้นเท่า ๆ กัน นั่นคือ การให้น้ำหนักของแต่ละจุดของเหตุการณ์ที่สนใจศึกษามีค่าเท่ากับ $1/n(\Omega)$ เมื่อ $n(\Omega)$ เป็นจำนวนผลลัพธ์ที่เป็นไปได้ทั้งหมด ดังนั้น ความน่าจะเป็นของเหตุการณ์ A ที่สนใจศึกษาในปริภูมิตัวอย่าง Ω เขียนแทนด้วย $P(A)$ ดังนี้

$$P(A) = \frac{\text{จำนวนผลลัพธ์ในเหตุการณ์ } A}{\text{จำนวนผลลัพธ์ที่เป็นไปได้ทั้งหมด}} = \frac{n(A)}{n(\Omega)}$$

ตัวอย่างที่ 1.4 กล่องใบหนึ่งบรรจุลูกบอล 8 ลูก สีแดง 3 ลูก และสีดำ 5 ลูก จงหาความน่าจะเป็น

1) สุ่มหยิบลูกบอล 1 ลูก และได้สีแดง

2) สุ่มหยิบลูกบอล 4 ลูกพร้อมกันและได้สีแดง 2 ลูก

วิธีทำ

1) จงหาความน่าจะเป็นสุ่มหยิบลูกบอล 1 ลูก และได้สีแดง

กำหนดให้ $n(\Omega)$ แทนจำนวนผลลัพธ์ที่เป็นไปได้ทั้งหมด นั่นคือ ลูกบอลมี 8 ลูกสุ่มมา 1 ลูก จะได้ $n(\Omega) = 8$ และ A แทนจำนวนผลลัพธ์ที่ได้สีแดง นั่นคือ $n(A) = 3$

ดังนั้น $P(A) = \dfrac{n(A)}{n(\Omega)} = \dfrac{3}{8} = 0.3750$

2) จงหาความน่าจะเป็นสุ่มหยิบลูกบอล 4 ลูกพร้อมกัน และได้สีแดง 2 ลูก

กำหนดให้ $n(\Omega)$ แทนจำนวนผลลัพธ์ที่เป็นไปได้ทั้งหมดในการสุ่มบอล 4 ลูกพร้อมกันจากบอล 8 ลูก โดยไม่คำนึงถึงอันดับในการจัด นั่นคือ $n(\Omega) = (8 \times 7 \times 6 \times 5)/(4 \times 3 \times 2 \times 1) = 70$ และ A แทนจำนวนผลลัพธ์ที่ได้สีแดง 2 ลูก โดยไม่คำนึงถึงอันดับในการจัด นั่นคือ $n(A) = [(3 \times 2 \times 1)/2] \times [(5 \times 4)/2] = 30$

ดังนั้น $P(A) = \dfrac{n(A)}{n(\Omega)} = \dfrac{30}{70} = 0.4286$

2) **ความน่าจะเป็นเชิงประจักษ์ (empirical probability)** หรือ ความถี่สัมพัทธ์ (relative frequency) คือ ความน่าจะเป็นของเหตุการณ์ที่สนใจศึกษาในปริภูมิตัวอย่าง Ω ของการทดลองสุ่มที่สนใจศึกษาจากการทำการทดลองซ้ำกันหลาย ๆ ครั้ง หรืออาศัยการเก็บรวบรวมข้อมูลในอดีตที่ผ่านมา n ครั้ง ความถี่สัมพัทธ์ของเหตุการณ์ A ที่สนใจศึกษา คือ

$$P(A) = \frac{\text{จำนวนผลลัพธ์ที่รวบรวมได้ของเหตุการณ์ที่สนใจศึกษา } A}{\text{จำนวนผลลัพธ์ที่รวบรวมทั้งหมด}} = \frac{n(A)}{n}$$

ตัวอย่างที่ 1.5 ร้านสะดวกซื้อแห่งหนึ่งต้องการสำรวจจำนวนลูกค้าที่เข้ามาซื้อของที่ร้าน ในแต่ละวันในช่วงระยะเวลา 4 สัปดาห์ บันทึกข้อมูลได้ดังนี้

วัน	อาทิตย์	จันทร์	อังคาร	พุธ	พฤหัสบดี	ศุกร์	เสาร์	รวม
จำนวนลูกค้า (คน)	102	45	50	60	68	250	108	683

จงหาความน่าจะเป็นที่จะมีลูกค้ามาซื้อของที่ร้านในวันศุกร์

2.2 ตัวแปรสุ่มและฟังก์ชันความน่าจะเป็น
(Random Variable and Probability Function)

นิยาม ตัวแปรสุ่ม (random variable, พจนานุกรมศัพท์สถิติศาสตร์ ฉบับราชบัณฑิตยสถาน, 2561 หน้า 273) คือ ตัวแปรที่ค่าต่าง ๆ ที่เกิดขึ้นถูกกำหนดด้วยกระบวนการเชิงความน่าจะเป็นหรือเชิงสุ่ม โดยเป็นฟังก์ชันค่าจริงที่มีโดเมนเป็นผลลัพธ์ที่กำหนดบนปริภูมิตัวอย่าง และมีเรนจ์เป็นจำนวนจริง สมมุติให้ X เป็นตัวแปรสุ่มบนปริภูมิเมเชอร์ $(\Omega,\mathcal{F})$ โดยที่ $\mathcal{F}$ เป็นซิกมาฟิลด์ของสับเซตของปริภูมิตัวอย่างที่สนใจศึกษา Ω และเรียกสมาชิกของ $\mathcal{F}$ ว่า เซตเมเชอร์เรเบิล (measurable set) บน Ω นั่นคือ

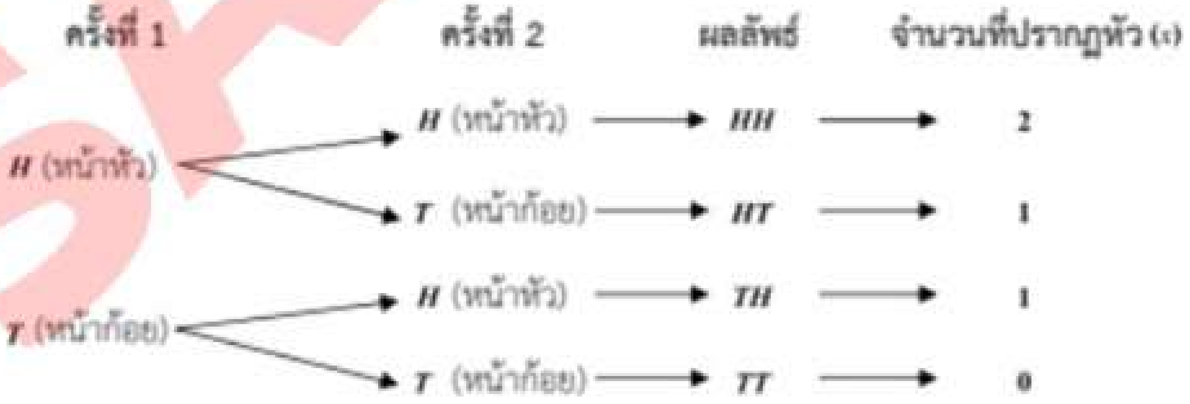
$X:\Omega \rightarrow \mathbb{R}$ ที่มีสมบัติ $\{\omega \in \Omega \mid X(\omega) \le x\} = \{X \le x\} \in \mathcal{F}$ สำหรับทุกค่าของ $x \in \mathbb{R}$ หรือ ถ้ามีอินเวอร์สภายใต้ X ของทุก ๆ เซตบอเรล (Borel set) B บน Ω (B เป็นซิกมาฟิลด์เล็กสุดเฉพาะกลุ่มของเซตที่ประกอบด้วยเซตเปิดทั้งหมดของ Ω) ใน $\mathbb{R}$ นั่นคือ

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F} \text{ สำหรับทุก ๆ } B \in \mathbb{R}$$

โดยทั่วไปจะสมมุติให้ $X(\omega)$ เป็นเซตของค่าที่เป็นไปได้ของตัวแปรสุ่ม X

$$(X = x) = \{\omega \in \Omega : X(\omega) = x\}$$

ตัวอย่างเช่น การทดลองโยนเหรียญเที่ยงตรง 1 เหรียญ จำนวน 2 ครั้ง หากผู้ทดลองสนใจความน่าจะเป็นที่เหรียญปรากฏหน้าหัว นั่นคือ ให้ $X(\omega)$ **เป็นจำนวนครั้งที่ปรากฏหัว** ω



จากปริภูมิตัวอย่าง คือ $\Omega = \{HH, HT, TH, TT\}$ ค่าของตัวแปรสุ่ม X คือ $x = 0, 1, 2$ แล้ว

$$X^{-1}(-\infty, x] = \begin{cases} \varnothing, & x < 0, \\ \{TT\}, & 0 \le x < 1, \\ \{TT, HT, TH\}, & 1 \le x < 2, \\ \Omega, & 2 \le x. \end{cases}$$

ดังนั้น X เป็นตัวแปรสุ่ม

และถ้านักวิจัยท่านหนึ่งสนใจศึกษาค่าใช้จ่ายใช้จ่ายในการใช้โทรศัพท์มือถือต่อเดือนของนักศึกษามหาวิทยาลัยแห่งหนึ่ง จะกำหนดให้ตัวแปรสุ่ม Y คือค่าใช้จ่ายในการใช้โทรศัพท์มือถือต่อเดือนของนักศึกษามหาวิทยาลัยแห่งนี้ ดังนั้น ปริภูมิตัวอย่าง คือ $\Omega = [0, +\infty)$ และค่าของตัวแปรสุ่ม Y ที่เป็นไปได้ คือ $y \ge 0$

$$Y(\omega) = \omega, \qquad \omega \in [0, +\infty)$$

นั่นคือ สามารถจำแนกตัวแปรสุ่มออกเป็น 2 ชนิด ตามชนิดของปริภูมิตัวอย่างที่ได้กล่าวไว้ในหัวข้อ 1.3 ดังนี้

1) ตัวแปรสุ่มไม่ต่อเนื่อง (discrete random variable) คือ ตัวแปรสุ่มที่มีปริภูมิตัวอย่างที่สามารถนับผลลัพธ์ทั้งหมดได้ ที่ประกอบด้วยจำนวนนับในรูปแบบจำกัดหรืออนันต์ ตัวอย่างเช่น จำนวนลูกค้าที่มาใช้บริการผ่านเคาน์เตอร์ของธนาคารแห่งหนึ่งในช่วงเวลาทำการ หรือจำนวนโทรศัพท์ที่โทรเข้ามาร้องทุกข์หรือแจ้งเหตุการณ์ต่าง ๆ ต่อวันของสถานีตำรวจแห่งหนึ่ง เป็นต้น

2) ตัวแปรสุ่มต่อเนื่อง (continuous random variable) คือ ตัวแปรที่มีปริภูมิตัวอย่างที่ไม่สามารถนับผลลัพธ์ทั้งหมดได้ นั่นคือ ตัวแปรมีความต่อเนื่องกันตลอดข้อมูล ซึ่งการวัดสามารถเป็นจุดทศนิยมได้ ตัวอย่างเช่น ค่าใช้จ่ายโทรศัพท์ต่อเดือน น้ำหนักของเด็กแรกเกิด เป็นต้น

จากสมบัติของการวัดความน่าจะเป็นในหัวข้อที่ 1.3 ที่ผ่านมา การอนุมานเชิงสถิติจะได้ประยุกต์ใช้การวัดความน่าจะเป็นของเหตุการณ์ที่สนใจศึกษา A ในปริภูมิตัวอย่าง Ω ให้อยู่ในรูปของฟังก์ชันความน่าจะเป็นของตัวแปรสุ่มที่สนใจศึกษา X นั่นคือ

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}, \quad A \in \Omega$$

โดยที่ $P(A) \ge 0$ เนื่องจาก $\mu(A) \ge 0$ และ $0 < \mu(\Omega) < \infty$

สมมุติให้ฟังก์ชัน $g : \Omega \to [0, +\infty)$ บนการวัด $\mu(A) = \sum_{x \in A} g(x), \ \ A \subseteq \Omega$

แล้ว

$$P(A)=\frac{\mu(A)}{\mu(\Omega)}=\frac{\sum_{x\in A}g(x)}{\sum_{x\in\Omega}g(x)},\quad A\subseteq\Omega$$

เรียกฟังก์ชันการวัดความน่าจะเป็นนี้เป็นฟังก์ชันความน่าจะเป็นของตัวแปรสุ่มไม่ต่อเนื่อง

สมมุติให้ฟังก์ชัน $g:\Omega\to[0,\infty)$, $\Omega\subseteq\mathbb{R}^n$ บนการวัด $\mu(A)=\int_A g(x)dx,\ A\subseteq\Omega$

แล้ว

$$P(A)=\frac{\mu(A)}{\mu(\Omega)}=\frac{\int_A g(x)dx}{\int_\Omega g(x)dx},\quad A\subseteq\Omega$$

เรียกฟังก์ชันการวัดความน่าจะเป็นนี้เป็นฟังก์ชันความน่าจะเป็นของตัวแปรสุ่มต่อเนื่อง

ดังนั้น ฟังก์ชันเชิงคณิตศาสตร์ที่แสดงความน่าจะเป็นของตัวแปรสุ่ม เรียกว่า **ฟังก์ชันความน่าจะเป็น (probability function)** เช่น กำหนดให้ $f_X(x)$ แทนฟังก์ชันความน่าจะเป็นของตัวแปรสุ่ม X หากตัวแปรสุ่ม X เป็นตัวแปรสุ่มไม่ต่อเนื่อง ฟังก์ชันที่ให้ค่าความน่าจะเป็นของค่าที่เป็นไปได้แต่ละค่าของ x จะเรียกว่าฟังก์ชันมวลความน่าจะเป็น (probability mass function: pmf) แทนด้วย $f_X(x)=P(X=x)$ นั่นคือ X เป็นตัวแปรสุ่มไม่ต่อเนื่องจะมีเซตที่สามารถนับได้ $E\in\mathbb{R}$ โดยที่ $p(x_i)=P\left(\omega\in\Omega \middle| X(\omega)=x_i, x_i\in E\right)=P\left(X=x_i\right)$ แล้ว $0\le p(x_i)\le 1$ และ $\sum_{all\ x_i}p(x_i)=1$

ตัวอย่างเช่น การทดลองโยนเหรียญเที่ยงตรง 1 เหรียญ จำนวน 2 ครั้ง ที่ผู้ทดลองสนใจความน่าจะเป็นที่เหรียญปรากฏหน้าหัว นั่นคือ ให้ X เป็นตัวแปรสุ่มของจำนวนที่ปรากฏหัว สามารถสร้างฟังก์ชันมวลความน่าจะเป็น ดังนี้

x	0	1	2	**รวม**
$f_X(x)=P(X=x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$	$\sum_{all\ x_i}p(x_i)=1$

กรณีตัวแปรสุ่ม X เป็นตัวแปรสุ่มต่อเนื่อง (continuous random variable or absolutely) ฟังก์ชันความน่าจะเป็นของตัวแปรสุ่มต่อเนื่องจะเรียกฟังก์ชันความหนาแน่นความน่าจะเป็น (probability density function: pdf) แทนด้วย $f_X(x)$ ที่มีสมบัติที่สำคัญดังนี้ $0\le f_X(x)\le 1$ ทุกค่าของ x และ $\int_{-\infty} f_X(x)dx=1$

สถิติศาสตร์ เป็นวิทยาศาสตร์ของการเรียนรู้จากข้อมูล การอนุมานเชิงสถิติอาศัยสารสนเทศที่ได้จากตัวอย่างสุ่ม เพื่อสรุปเกี่ยวกับค่าพารามิเตอร์ที่ไม่ทราบค่าที่ผู้สนใจศึกษา มีข้อสมมุติเกี่ยวกับฟังก์ชัน การแจกแจงความน่าจะเป็นของสิ่งที่สนใจศึกษาไว้ หนังสือเล่มนี้มุ่งเน้น การนำความน่าจะเป็นและสมบัติที่สำคัญ ที่จำเป็นต่อการอนุมานสถิติขั้นสูงต่อไป ไม่ว่าจะเป็นการประมาณค่าพารามิเตอร์แบบจุดและแบบช่วง การตั้งสมมุติฐานและการทดสอบสมมุติฐาน พร้อมทั้งการเสนอแนวทางในการประยุกต์ใช้ความน่าจะเป็น ในการอนุมานเชิงสถิติของผู้เขียนเป็นตัวอย่างประกอบ และการใช้โปรแกรมอาร์ช่วยอธิบายปัญหา สามารถใช้หนังสือเล่มนี้เพื่อประกอบการศึกษาและการค้นคว้าเพิ่มเติมได้

CHIANG MAI
UNIVERSITY PRESS

ISBN (e-Book) : 978-616-398-527-9
9 786163 985279
ราคา 400 บาท