3. ตัวแปรเชิงสุ่มและการแจกแจงน่าจะเป็น

The most important Questions of Life are,
for the most part, really only Problems of
Probabilitys

Pierre Simon de Laplace.

การทดลองเชิงสุ่มบางประเภทให้ผลทดลองเป็นตัวเลข (numbers) ทันที การทดลองประเภทนี้มักจะมีการวัดหรือนับปริมาณบางอย่าง เช่น นับจำนวนหน่วยกิตรวมที่นักศึกษาปีที่สองสอบผ่าน ผลลัพธ์จากการนับ (ซึ่งเป็นตัวเลข) จะเป็นผลจากการทำการทดลองเชิงสุ่ม เพราะเราไม่ทราบล่วงหน้าก่อนนับว่าคนหนึ่งจะผ่านเท่าใด? ในกรณีนี้กลุ่มผลทดลองจะเป็น $S = \{0, 1, 2, \ldots\}$ ตัวเลขอาจจะได้มาจากการวัด เช่น วัดเวลาที่สัตว์ใช้เรียนรู้ การนับหรือวัดซึ่งเป็นการทดลองเชิงสุ่มทำนองนี้เราเรียกว่า **การนับหรือการวัดเชิงสุ่ม**

ก่อนที่จะทำการนับหรือวัดแล้วเสร็จ ตัวเลขที่จะได้เป็นผลลัพธ์นั้นเรามักจะแทนด้วย X (หรือตัวอื่น ๆ เช่น Y, E เป็นต้น) และเรียก X ว่า ตัวแปรเชิงสุ่ม จงสังเกตว่า X เป็นแต่เพียงสัญญลักษณ์แทนตัวเลขที่ยังไม่เป็นค่าใดค่าหนึ่งจนกว่าจะทำการนับหรือวัดเชิงสุ่มแล้วเสร็จ เมื่อทำการวัดหรือนับเชิงสุ่มเสร็จแล้วตัวแปรเชิงสุ่ม X จึงจะมีผลลัพธ์เป็นค่าใดค่าหนึ่งโดยเฉพาะ

ในการศึกษาถึงตัวแปรเชิงสุ่มนี้เรามิได้มุ่งสนใจว่า X จะบอกผลลัพธ์เป็นอะไร ความสนใจจะมุ่งอยู่ที่สถานะการณ์ก่อนจะทำการนับหรือวัดแล้วเสร็จและต้องการหาความน่าจะเป็นที่ X ออกผลลัพธ์มาเป็นค่าใดค่าหนึ่ง เช่น x หรือออกผลลัพธ์มาอยู่ในช่วงหนึ่งเช่นช่วง (a,b) เราจะแทนความน่าจะเป็นของมันด้วย $P(X = x)$ และ $P(a < X < b)$ ตามลำดับ

ในทางคณิตศาสตร์เรามองตัวแปรเชิงสุ่มแบบฟังก์ชั่นเชิงคณิตศาสตร์อย่างหนึ่ง โดยมีนิยามดังนี้

นิยาม ตัวแปรเชิงสุ่ม X เป็นฟังก์ชั่นค่าจริง (Real - valued function) ในตัวแบบน่าจะเป็น (S, P) โดยที่มันจะกำหนดแต่ละผลทดลองด้วยเลขจำนวนจริง ซึ่งเรียกว่าค่าของตัวแปรเชิงสุ่มจากผลทดลองนั้น

นั่นก็คือตัวแปรเชิงสุ่มเป็นฟังก์ชั่นค่าจริงที่มีโดเมนเป็นกลุ่มผลทดลองของการทดลองเชิงสุ่ง และมีนิสัยเป็นเซ้ทของเลขจำนวนจริง

ดังนั้นตัวแปรเชิงสุ่มจะให้ค่าหนึ่งต่อแต่ละผลทดลอง เช่น ถ้ากำหนดผลทดลอง s ของกลุ่มผลทดลอง S แล้ว $X(s)$ จะเป็นเลขจำนวนจริงจำนวนหนึ่ง ถ้าให้ X เป็นตัวแปรเชิงสุ่มที่มีค่าเป็น $x_1, x_2, \ldots,$ นั่นก็แสดงว่าพิสัยของ X เป็น $\{x_1, x_2 \ldots\}$ นั่นเอง สำหรับ $X = x_i$ $(i = 1, 2, 3, \ldots)$ นี้จะใช้แทนเหตุการณ์ที่ประกอบด้วยผลทดลองที่แปลงไปเป็น x_i ด้วย X นั่นคือ

$$(X = x_i) = \{s \leq S \mid X(s) = x_i\}$$

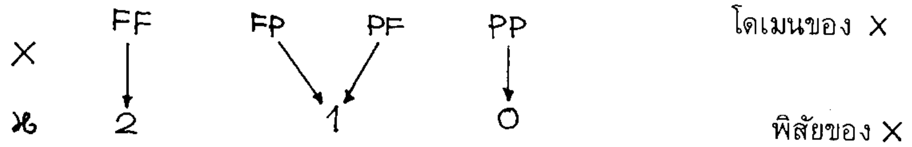
และเราจะเขียนความน่าจะเป็นของเหตุการณ์ $X = x_i$ ด้วย $f(x_i)$ หรือ $P(x_i)$ซึ่งจะมีความหมายเท่ากับ $P(X = x_i)$

ตัวอย่าง ในการทดลองโดยการสุ่มวิชาที่ลงทะเบียนมาสองวิชา เมื่อให้ X เป็นตัวแปรเชิงสุ่มซึ่งแทนจำนวนวิชาที่สอบไม่ผ่าน (F) แล้วกลุ่มผลทดลองจะเป็นดังนี้

$$S = \{FF, FP, PF, PP\};$$ ในเมื่อ P เป็นวิชาที่สอบผ่าน

และกลุ่มผลทดลองนี้จะเป็นโดเมนของ X สำหรับของค่าที่แทนจำนวนวิชาที่สอบไม่ผ่าน $\{0, 1, 2\}$ จะเป็นพิสัยของ X นั่นคือเป็นเซ็ทของค่าที่ X จะมีได้

ตัวแปรเชิงสุ่ม X จะแปลงแต่ละจุดหรือผลทดลองในโดเมนไปยังจุดหนึ่งในพิสัยดังนี้



บางทีเราเขียนเป็นแบบคณิตศาสตร์ได้ดังนี้

$X(FF) = 2$ $X(PF) = 1$

$X(FP) = 1$ $X(PP) = 0$

ซึ่งหมายความว่าถ้าผลทดลองเป็น FF แล้วให้ X มีค่าเป็น 2 ถ้าเป็น FP หรือ PF ให้เป็น 1, และเป็น PP ให้ 0

3.1 ประเภทของตัวแปรเชิงสุ่ม (Classification of Random Variables)

เนื่องจากตัวแปรเชิงสุ่มเป็นฟังก์ชั่นที่นิยามหรือกำหนดในกลุ่มผลทดลอง และโดยที่กลุ่มผลทดลองมีอยู่ 2 ประเภท เราจึงแบ่งประเภทของตัวแปรเชิงสุ่มตามกลุ่มผลทดลองได้เป็น 2 ประเภทเหมือนกัน ดังนี้

(1) ตัวแปรเชิงสุ่มไม่ต่อเนื่อง (Discrete Random Variable) เป็นตัวแปรเชิงสุ่มที่กำหนดในกลุ่มผลทดลองแบบไม่ต่อเนื่อง ตัวแปรเชิงสุ่มประเภทนี้มักจะเกิดจากการนับเชิงสุ่ม ตัวอย่าง เช่น จำนวนนักศึกษาที่สอบผ่านแต่ละวิชา ซึ่งเป็นวิชาบังคับของนักศึกษาจิตวิทยาจะเป็นตัวแปรเชิงสุ่มไม่ต่อเนื่องที่มีค่าหนึ่งค่าใดในเซ็ทของ $0, 1, 2, \ldots$ นั่นคือผลลัพธ์ของตัวแปรเชิงสุ่มจะเป็นเลขจำนวนเต็ม แต่ในกรณีอื่นถึงแม้ว่าจะเป็นตัวแปรเชิงสุ่มไม่ต่อเนื่องผลลัพธ์อาจจะไม่เป็นเลขจำนวนเต็มก็ได้ เช่น กรณีที่กลุ่มผลทดลองของ X มีตัวเลขที่เป็นไปได้เพียง 5 จำนวนคือ

$$S = \{3.50 , 3.78 , 15.12 , 17.25 , 25.30\}$$

X ก็ยังคงเป็นตัวแปรเชิงสุ่มไม่ต่อเนื่อง เพราะกลุ่มผลทดลองเป็นแบบไม่ต่อเนื่อง กรณีนี้ผลลัพธ์ของ X ไม่เป็นเลขนำนวนเต็ม

(2) ตัวแปรงเชิงสุ่มต่อเนื่อง (Continuous Random Variable) เป็นตัวแปรเชิงสุ่มที่นิยามในกลุ่มผลทดลองแบบต่อเนื่อง เช่น เวลาที่เด็กใช้เรียนรู้การบวกที่มีผลลัพธ์ไม่เกิน 10 จะเป็นตัวแปรเชิงสุ่มต่อเนื่อง ซึ่งจะมีค่าเป็นค่าหนึ่งค่าใดในช่วง (a,b) ในเมื่อ a และ b เป็นเลขจำนวนจริงที่ $a < b$

ถ้าจะพิจารณาในค่าของตัวแปรเชิงสุ่มหรือกฎในการกำหนดเลขจำนวนจริงให้แก่ผลทดลองในกลุ่มผลทดลองแล้วเราสามารถแยกออกตัวแปรเชิงสุ่ม ออกเป็น 4 ประเภทตามระดับของการวัด (Level of Measmement) โดยที่การวัดเป็นการกำหนดตัวเลขให้แก่สิ่งของหรือเหตุการณ์ตามกฎต่าง ๆ และกฎต่าง ๆ กันนี้เองจะให้ระดับของการวัดต่าง ๆ กันด้วย

(1) ตัวแปรเชิงสุ่มแบบนามบัญญัติ (Nominal or Categorical Random Variables) ค่าของตัวแปรเชิงสุ่มประเภทนี้ใช้ขนานชื่อให้กับผลทดลองหรือเหตุการณ์ต่าง ๆ เช่น หญิง ชาย เป็นต้น ตัวเลขที่เป็นค่าของตัวแปรเชิงสุ่มประเภทนี้จะมีคุณค่า (value) เหมือนกัน และอาจใช้สลับสับเปลี่ยนหรือทดแทนกันได้ เช่น 0 ให้เป็นชาย และ 1 เป็นหญิง หรือจะให้ 1 เป็นชาย และ 0 เป็นหญิงก็ได้

(2) ตัวแปรเชิงสุ่มแบบเรียงอันดับ (Ordinal or Rank Random Variables) ค่าของตัวแปรเชิงสุ่มประเภทนี้ใช้เรียง หรือจัดอันดับผลทดลองในสกุลเดียวกันให้ลดหลั่นกันเป็นขั้น ๆ ตามปริมาณและคุณภาพมากน้อย เช่นระดับรายได้ต่ำกว่า, และสูง ก็กำหนดตัวเลขเป็น 1, 2, 3 เป็นต้น

(3) ตัวแปรเชิงสุ่มแบบอันตรภาค (Interval Random Variables) ค่าของตัวแปรเชิงสุ่มประเภทนี้ใช้แบ่งตัวกำหนด ผลทดลองในสกุลเดียวกันออกเป็นประเภท ๆ หรือเป็นจังหวะเป็นช่วง ๆ ขั้น ๆ ที่มีขนาดใหญ่เท่า ๆ กัน อย่างไรก็ตามตัวเลขที่กำหนดนั้นถ้าเป็น 0 ก็ไม่ได้

หมายความเป็นศูนย์แท้จริงศูนย์อนันต์ (Absolute zero) แต่เป็นศูนย์นิยมหรือศูนย์สัมพัทธ์ (Relative of arbitrary zero) เท่านั้น และจำนวนเท่าของค่าของตัวแปรเชิงสุ่มจะไม่เป็นจริง เช่นค่าของตัวแปรที่เป็นคะแนนสอบ ถ้าได้ 0 ก็ไม่ได้หมายความว่าไม่มีความรู้ หรือได้ 50 ก็ไม่ใช่มีความรู้เป็น 2 เท่าของ 25 เป็นต้น

(4) ถ้าแปรเชิงสุ่มแบบอัตรามส่วน (Ratio Random Variables) ค่าของตัวแปรประเภทนี้เป็นจำนวนเลขที่สมบูรณ์ที่สุด นั่นคือมีศูย์แท้จริงศูนย์อนันต์ มีหน่วยเป็นขนาดโตเท่า ๆ กัน และเรียงขึ้นลงตามลำดับสม่ำเสมอกัน และจำนวนเท่าของค่าจะเป็นจริง เช่นค่าของแปรเชิงสุ่มที่ใช้แทนน้ำ**หนัก,** ความยาว, ความหนาแน่น เป็นต้น

	ลักษณะที่สำคัญ		
ตัวแปรเชิงสุ่ม	ก. อันดับ	ข. ระยะทาง	ค. จุดกำเนิด
นามบัญญัติ	×	×	×
เรียงอันดับ	✓	×	×
อันตรภาค	✓	✓	×
อัตราส่วน	✓	✓	✓

ก. อันดับ (Sequence or Order) ของจำนวนมีความหมายหรือไม่?

ข. ระยะทาง (Distance) ระหว่างสองจำนวนมีความหมายหรือไม่?

ค. จุดกำเนิด (Origin) มีความหมายเดียว (Unique) หรือไม่?

ตัวอย่าง ในสัปดาห์หนึ่งเราอาจจะใช้เครื่องมือสำหรับเขียนหลายชนิด เช่น ดินสอ ปากกาหมึกซึม ปากกาลูกลื่น ปากกาแมจิก เป็นต้น ตัวแปรเชิงสุ่มแบบใดที่จะใช้เหตุการณ์ต่อไปนี้

1. ให้เบอร์เครื่องเขียน
2. เรียงลำดับเครื่องเขียนตามความชอบ
3. วัดอุณหภูมิต่ำสุดที่เครื่องเขียนจะใช้งานได้
4. วัดน้ำหนักเครื่องเขียน

ตัวแปรเชิงสุ่มที่ใช้ในแต่ละเหตุการณ์จะต่างกัน นั่นคือเป็นตัวแปรเชิงสุ่มแบบนามบัญญัติ เรียงอันดับ อันตรภาค และอัตราส่วน

3.2 ฟังก์ชั่นมวลน่าจะเป็น (Probability Mass Function, PMF)

ตัวแปรเชิงสุ่มแบบไม่ต่อเนื่องมีลักษณะที่สำคัญดังนี้

(1) ค่าของตัวแปรที่เป็นไปได้

(2) ความน่าจะเป็นที่เกี่ยวพัน (associated) กับแต่ละค่าที่เป็นไปได้จากการที่เราทราบว่าความน่าจะเป็นของแต่ละผลทดลองในกลุ่มผลทดลองนั้นเป็นเท่าใด แล้วเราก็สามารถกำหนดความน่าจะเป็นที่ตัวแปรเชิงสุ่มจะมีค่าเป็นค่าใด ได้กฎหรือวิธีการสำหรับกำหนดความน่าจะเป็นให้แก่ค่าของตัวแปรเชิงสุ่มแบบไม่ต่อเนื่องนั้นจะเรียกว่าฟังก์ชันมวลน่าจะเป็น P.M.F. **หรือการแจกแจงน่าจะเป็น** (Probability Distribution) ถ้าเป็นไปตามนิยามต่อไปนี้

นิยาม กำหนดให้ X เป็นตัวแปรเชิงสุ่มแบบไม่ต่อเนื่องที่มีค่าเป็น $x_1, x_2, x_3 \ldots$ ...... ฟังก์ชัน f หรือ P จะเรียกว่าฟังก์ชันมวลน่าจะเป็นของ X ถ้า f กำหนดไว้ดังนี้

$$f(x_i) = P(X = x_i); \quad i = 1, 2, 3, \ldots\ldots$$

และ $f(x_i)$ จะต้องมีคุณสมบัติดังนี้

$$0 \leq f(x_i) \leq 1$$
$$\sum_i f(x_i) = 1$$

ตัวอย่าง ถ้ากระทำการทดลองอย่างหนึ่งโดยการเดาข้อทดสอบแบบถูกผิด จำนวน 3 ข้อ และให้ X เป็นตัวแปรเชิงสุ่มที่แทนจำนวนข้อที่ถูก แล้วเราจะได้ผลทดลองที่เป็นไปได้ทั้งหมด ความน่าจะเป็นของแต่ละผลทดลอง, และค่าของตัวแปรเชิงสุ่ม X ดังตารางต่อไปนี้ (R แทนเดาถูก และ W แทนเดาผิด)

ผลทดลอง (O_i)	RRR	RRW	RWR	WRR	RWW	WRW	WWR	WWW
ความน่าจะเป็น, $P(O_i)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
ค่าของ X, x_i	3	2	2	2	1	1	1	0

สำหรับค่าที่เป็นไปได้ของตัวแปรเชิงสุ่ม X กับความน่าจะเป็นที่กำกับอยู่ดังตารางต่อไปนี้จะเป็นการแจกแจงน่าจะเป็นของตัวแปรเชิงสุ่ม X

x_i	0	1	2	3
$f(x_i)$	1/8	3/8	3/8	1/8

สำหรับการแจกแจงน่าจะเป็นนั้นเราสามารถเขียนให้อยู่ในรูปเซ็ท, สมการ, หรือกราฟก็ได้ ดังนี้

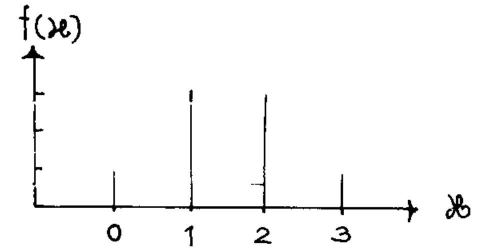
(1) เซ็ท $f = \{(0, 1/8), (1, 3/8), (2, 3/8), (1, 1/8)\}$

(2) สูตรหรือสมการ

$$f(x) = 1/8 \ , \quad x = 0,3$$
$$\quad\;\; = 3/8 \ , \quad x = 1,2$$

หรือ $f(x) = \binom{3}{x} (1/2)^{x} (1 - 1/2)^{3-x} \ ; \ x = 0,1,2,3$

(3) กราฟ



ตัวอย่าง ในการศึกษาผลการสอบวิชาจิตวิทยาการเมือง สมมติว่าผลการสอบที่จะเป็นไปได้เป็น A, B, C, D และ F เมื่อให้ X แทนคะแนนที่เป็นผลของการสอบ แล้ว X จะมีค่าที่เป็นไปได้ดังนี้

ผลการสอบ	A	B	C	D	F
x_1	4	3	2	1	0

สำหรับความน่าจะเป็นที่กำหนดให้แก่แต่ละเหตุการณ์นั้นจะต้องเป็นความน่าจะเป็นเชิงอัตนัย ซึ่งสมมติว่าเป็นดังนี้

x	0	1	2	3	4
$f(x)$	.1	.1	.2	.4	.2

3.3 ฟังก์ชั่นหนาแน่นน่าจะเป็น (Probability Density Function, PDF)

ในกรณีที่ชุดของตัวเลขที่เป็นพิสัยของตัวแปรเชิงสุ่มมีจำนวนมากเป็นอนันต์นับไม่ได้ (Uncountably infinite) นั่นก็คือกรณีของตัวแปรเชิงสุ่มต่อเนื่องนั่นเอง สำหรับกลุ่มผลทดลองแบบไม่ต่อเนื่องนั้นเราสามารถกำหนดความน่าจะเป็นที่ไม่เป็นศูนย์ให้แก่แต่ละค่าของตัวแปรเชิงสุ่มได้ถึงแม้ว่าค่าของตัวแปรเชิงสุ่มจะเป็นอนันต์นับได้ (Countable infinite) และความน่าจะเป็นที่กำหนดให้นั้นเมื่อรวมกันจะเท่ากับหนึ่งพอดี แต่ในกลุ่มผลทดลองแบบต่อเนื่องนี้จะเป็นไปไม่ได้ เพราะถ้าพยายามสร้างเหตุการณ์แบบง่าย ๆ ที่ประกอบด้วยผลทดลองเพียงหนึ่งและกำหนด

ความน่าจะเป็นให้แก่แต่ละเหตุการณ์ง่าย ๆ เหล่านี้เพื่อที่จะให้ผลรวมของความน่าจะเป็นให้กับหนึ่งแล้วเราก็ไม่สามารถทำได้ ดังนั้นจึงหันมาสนใจเหตุการณ์ผสมที่ประกอบด้วยผลทดลองมาก ๆ นั่นคือสนใจเหตุการณ์ที่ตัวแปรเชิงสุ่มที่มีค่าอยู่ในช่วงใดช่วงหนึ่ง แล้วเราจะได้นิยามต่อไปนี้

นิยาม ให้ X เป็นตัวแปรเชิงสุ่มแบบต่อเนื่องที่มีพิสัยเป็นเซ็ทของเลขจำนวนจริงที่อยู่ในช่วง (a, b), $a < b$ ฟังก์ชั่น f จะเรียกว่าฟังก์ชั่นหนาแน่นน่าจะเป็น ถ้า $f(x)$ สอดคล้องกับคุณสมบัติต่อไปนี้

– $f(x) \geq 0 \qquad a \leq x \leq b$

– $\int_{-\infty}^{\infty} f(x)\,dx = 1$

– ความน่าจะเป็นที่ X มีค่าอยู่ในช่วง $[x, x + dx]$ จะเท่ากับ $f(x)\,dx$

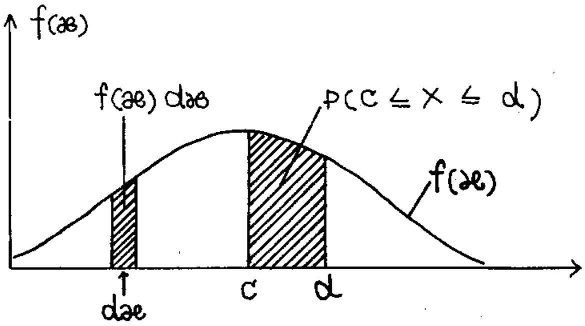
นั่นคือ $P(x \leq X \leq x + dx) = f(x)\,dx$

สำหรับความน่าจะเป็นที่ X มีค่าอยู่ในช่วง $[x, x, + dx]$ หรือ $f(x)\,dx$ นั้นก็คือพื้นที่ภายใต้ฟังก์ชั่นหรือโค้ง $f(x)$ ระหว่างจุด x กับ $x + dx$ ดังนั้นความน่าจะเป็นที่ X จะมีค่าอยู่ในช่วง (c, d) จึงสามารถกำหนดได้ดังนี้

$$P[c \leq x \leq d] = \int_{c}^{d} f(x)\,dx$$

ซึ่งเท่ากับพื้นที่ภายใต้ฟังก์ชั่น $f(x)$ ระหว่างจุด c กับ d นั่นเอง

ฟังก์ชั่นหนาแน่นน่าจะเป็นสามารถแสดงได้ดังกราฟต่อเนื่อง ต่อไปนี้



ตัวอย่าง ตัดเหล็กแท่งให้มีความยาวแบบสุ่ม ซึ่งไม่เกิน 12 นิ้ว ฟังก์ชั่นหนาแน่นน่าจะเป็นสำหรับตัวแปรเชิงสุ่ม ที่แทนความยาวของเหล็กแท่งซึ่งเป็นนิ้วจะหาได้ดังนี้

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{12} a\,dx = 1$$

$$ax\Big|_{0}^{12} = 1 \quad , \quad a = 1/12$$

ดังนั้น $$f(x) = 1/12 \;, \quad 0 < x < 12$$

ถ้าเราจะเสนอฟังก์ชั่นหนาแน่นน่าจะเป็นในรูปตาราง เราต้องให้ค่าของตัวแปรเชิงสุ่มเป็นช่วง ๆ ดังนี้

x	$f(x)$
$0 < x < 4$	$1/3$
$4 < x < 8$	$1/3$
$8 < x < 12$	$1/3$

3.4 ฟังก์ชั่นแจกแจงสะสม (Cumulative Distribution Function, CDF)

ตัวแปรเชิงสุ่มนอกจากจะเสนอหรือบรรยายคุณลักษณะของมันด้วย **PMF** หรือ **PDF** แล้วเรายังนิยมบรรยายคุณลักษณะของตัวแปรเชิงสุ่มด้วย ฟังก์ชั่นแจกแจงสะสมดังนิยามต่อไปนี้

นิยาม ฟังก์ชั่น F จะเรียกว่าฟังก์ชั่นแจกแจงสะสมของตัวแปรเชิงสุ่ม X ถ้า F กำหนดไว้ดังนี้

$$F(x) = P(X \le x) = \begin{cases} \displaystyle\sum_{X \le x} f(x) & \text{ตัวแปรไม่ต่อเนื่อง} \\ \displaystyle\int_{-\infty} f(x)\,dx & \text{ตัวแปรต่อเนื่อง} \end{cases}$$

และ $F(x)$ ต้องมีคุณสมบัติดังต่อไปนี้

(1) $F(x)$ จะเป็นฟังก์ชั่นที่มีค่าเพิ่มขึ้นเมื่อ x มีค่ามากขึ้น ซึ่งเราเรียกว่า**ฟังก์ชั่นขยาย** (Nondecreasing function) นั่นคือ ถ้า $x_1 < x_2$ แล้ว $F(x_1) < F(x_2)$

(2) $F(+\infty) = 1$ และ $F(-\infty) = 0$ หรือ $\lim\limits_{x \to -\infty} F(x) = 0$ และ $\lim\limits_{x \to +\infty} F(x) = 1$ นั่นก็คือ $0 \le F(x) \le 1$

(3) ถ้า X เป็นตัวแปรเชิงสุ่มแบบต่อเนื่องแล้ว $f(x) = \frac{dF(x)}{dx}$ และถ้า X เป็นตัวแปรเชิงสุ่มแบบไม่ต่อเนื่องที่มีค่าเป็น $x_1, x_2, \cdots, x_n$ โดยที่ $x_1 \angle x_2 \angle x_3 \angle \ldots \ldots \angle x_n$ แล้ว $f(x_i) \quad F(x_i) \quad .F(x_{i-1})$

(4) ถ้า X เป็นตัวแปรเชิงสุ่มแบบไม่ต่อเนื่อง แล้ว $F(X)$ จะเป็นฟังก์ชั่นแบบขั้นบันได (Step function) ซึ่งเป็นฟังก์ชั่นที่ต่อเนื่องทางขวา (Continuous on the right) แต่ไม่ต่อเนื่องทางซ้าย และขั้นบันไดนั้นจะเกิดขึ้นที่จุด $X = x_i$ พอดี

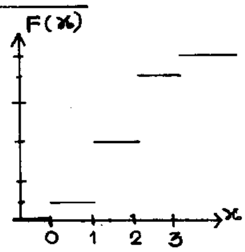
ถ้า X เป็นตัวแปรเชิงสุ่มแบบต่อเนื่อง แล้ว $F(X)$ จะเป็นฟังก์ชั่นที่ต่อเนื่องด้วย

ตัวอย่าง ให้ X เป็นตัวแปรเชิงสุ่มที่มีฟังก์ชั่นมวลน่าจะเป็นดังนี้

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

เราสามารถหาฟังก์ชั่นแจกแจงสะสม $F(x)$ ในรูปของตารางหรือกราฟได้ดังนี้

ค่าของตัวแปรเชิงสุ่ม x	ความน่าจะเป็น $F(x)$
$x \angle 0$	0
$x \leq 0$	1/8
$x \leq 1$	4/8
$x \leq 2$	7/8
$x \leq 3$	1



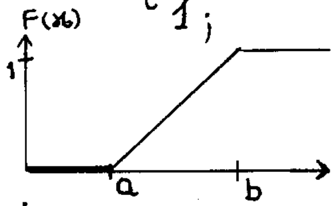
ตัวอย่าง ให้ X เป็นตัวแปรเชิงสุ่มที่มีฟังก์ชั่นหนาแน่นน่าจะเป็นดังนี้

$$f(x) = 1/(b-a); \quad a \angle x \angle b$$

ดังนั้นฟังก์ชั่นแจกแจงสะสมของ X คือ

$$f(x) = \begin{cases} 0; & \\ (x-a)/(b-a); & a \leq x \angle b \\ 1; & x \geq b \end{cases}$$

กราฟของมันก็คือ



3.5 มาตรฐานวัดสรุปของตัวแปรเชิงสุ่ม (Summary Measures of Random Variable)

เราได้พิจารณาคุณลักษณะของตัวแปรเชิงสุ่มในแง่ฟังก์ชั่นน่าจะเป็นและฟังก์ชั่นแจกแจงสะสมมาแล้ว ต่อไปเราจะพิจารณาคุณลักษณะของตัวแปรเชิงสุ่มโดยสรุปด้วยดัชนีหรือ

ปริมาณที่เป็นเลขจำนวนเดี่ยว ๆ ปริมาณนี้เราเรียกว่าพารามิเตอร์ (Parameter) ของการแจกแจงตัวแปรเชิงสุ่ม มาตราวัดสรุปของตัวแปรเชิงสุ่มที่สำคัญมีดังนี้

- มาตราวัดค่าเฉลี่ย และมาตรา วัดตำแหน่ง
- มาตราวัดการกระจาย

3.5.1 **มาตราวัดค่าเฉลี่ย** (Measures of Location) เป็นดัชนีหรือปริมาณจำนวนหนึ่งซึ่งแสดงคุณลักษณะของตัวแปรเชิงสุ่มด้วยค่าเฉลี่ยค่ากลาง ๆ (Expected, typical or Normative value) เรามีวิธีวัดค่าเฉลี่ยแบบต่าง ๆ ดังนี้

ก. **กึ่งพิสัย** (Midrange) เป็นมาตราวัดที่หยาบที่สุด แต่หาได้รวดเร็ว ดังนิยามต่อไปนี้

นิยาม ถ้า x_i เป็นค่าของตัวแปรเชิงสุ่ม X แล้วกึ่งพิสัย, μ_r จะกำหนดไว้ว่า

$$\mu_r = \frac{1}{2}[\max(x_i) + \min(x_i)]$$

ในเมื่อ $\max(x_i)$ และ $\min(x_i)$ เป็นค่าของตัวแปรเชิงสุ่มที่มีค่ามากสุด และน้อยสุดตามลำดับ

ตัวอย่าง ถ้า X มีการแจกแจงน่าจะเป็นดังนี้

x	1	2	3	4
$f(x)$	.1	.4	.3	.2

แล้ว

$$\mu_r = \frac{1}{2}(4+1) = 2.5$$

ข. **ฐานนิยม** (Mode)

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่มแล้ว μ_o จะเป็นฐานนิยมของ X ถ้า μ_o เป็นค่าที่ทำให้ฟังก์ชั่นของ X หรือ $f(x)$ มีค่าสูงสุด

นั่นคือถ้า X เป็นตัวแปรเชิงสุ่มแบบไม่ต่อเนื่อง ค่าของ X ที่มีความน่าจะเป็น $f(x) = P(X=x)$ สูงสุดจะเป็นฐานนิยม

และถ้า X เป็นตัวแปรเชิงสุ่มแบบต่อเนื่อง μ_o จะเป็นฐานนิยมของ X ถ้าสอดคล้องกับสมการ

(1) $f'(x)\big|_{x=\mu_o} = 0$

และ (2) $f''(x)\big|_{x=\mu_o} < 0$

ฐานนิยมของตัวแปรเชิงสุ่มจะหาได้เสมอ (Exist) และไม่จำเป็นต้องมีตัวเดียว แต่ถ้ามีหลายตัวก็ไม่มีประโยชน์ในการเป็นมาตราวัดสรุป

เราสนใจฐานนิยมเมื่อต้องการทราบว่าเหตุการณ์ไหนหรือค่าไหนของตัวแปรเชิงสุ่ม

มีโอกาสเกิดขึ้นบ่อยที่สุด เช่นสนใจว่าจำนวนเด็กในครอบครัวเป็นอย่างไร แล้วเราใช้ฐานนิยมดีกว่าอย่างอื่น

ตัวอย่าง ตัวแปรเชิงสุ่ม x มีการแจกแจงน่าจะเป็นดังนี้

x	1	2	3
$f(x)$	.5	.2	.3

เนื่องจาก $f(1)$ มีค่าสูงสุดคือ .5 ดังนั้น $\mu_0 = 1$ จึงเป็นฐานนิยมซึ่งใช้วัดค่าเฉลี่ยได้ดีพอสมควร

ตัวอย่าง ถ้าตัวแปรเชิงสุ่ม X มีการแจกแจงดังนี้

$$f(x) = e^{-x} \quad ; \quad x > 0$$

เนื่องจาก $f(0)$ มีค่าสูงสุด ดังนั้น μ_0 จึงเป็นฐานนิยม แต่ใช้วัดค่าเฉลี่ยได้ไม่ดีเลย

ตัวอย่าง ถ้าตัวแปรเชิงสุ่ม X มีการแจกแจงดังนี้

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad ; \; -\infty < x < \infty$$

เนื่องจาก $f'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}(-x) = 0$ เราจะได้ $X = +\infty, 0, -\infty$
แต่ $f''(0) < 0$ ดังนั้น $\mu_0 = 0$ เป็นฐานนิยมซึ่งใช้วัดค่าเฉลี่ยได้ดี

ค. **มัธยฐาน** (Median) เป็นค่าของตัวแปรเชิงสุ่มที่แบ่งการแจกแจงของตัวแปรเชิงสุ่มนั้นออกเป็นสองส่วนเท่า ๆ กัน ดังนิยาม

นิยาม ถ้า μ_m เป็นค่าของตัวแปรเชิงสุ่มแบบไม่ต่อเนื่อง X และทำให้

$$P(X \leq \mu_m) \geq 1/2 \quad \text{และ} \quad P(X \geq \mu_m) \geq 1/2$$

แล้ว μ_m จะเรียกว่ามัธยฐานของตัวแปรเชิงสุ่ม X

แต่ถ้า X เป็นตัวแปรเชิงสุ่มแบบต่อเนื่อง แล้วมัธยฐาน μ_m ของ X จะทำให้ $F(\mu_m) = P(X \leq \mu_m) = 1/2$ พอดี และเราสามารถหา μ_m ได้จากสมการ

$$\int_{-\infty}^{\mu_m} f(x)\,dx = 1/2$$

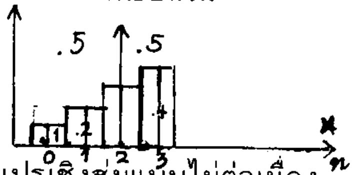
ตัวอย่าง กำหนดฟังก์ชันน่าจะเป็นของตัวแปรเชิงสุ่ม X ดังนี้

x	0	1	2	3
$f(x)$	.1	.2	.3	.4

$P(X \geq 2) = 0.7 > 0.5$ $\mu_m = 2$ เป็นมัธยฐาน เพราะ $P(X \leq 2) = 0.6 > 0.5$ และ

มัธยฐานของตัวแปรเชิงสุ่มสี่เราสามารถหาได้เที่ยงตรงได้โดยการเทียบดังนี้

x	0	1	2	3
$F(x)$	.1	.3	.6	1.0



$$\mu_m = 1.5 + \frac{0.5 - 0.3}{0.6 - 0.3} = 2.17$$

จากนิยามและตัวอย่างเราจะเห็นได้ว่ามัธยฐานของตัวแปรเชิงสุ่มแบบไม่ต่อเนื่องจะมีเสมอ แต่ไม่จำเป็นว่าต้องมีตัวเดียว

ตัวอย่าง กำหนดฟังก์ชั่นหนาแน่นน่าจะเป็นของตัวแปรเชิงสุ่มไว้ดังนี้

$$f(x) = 2x/3 \; ; \; 1 < x < 2$$

เราหามัธยฐาน μ_m ได้จากสมการ

$$\int_{-\infty}^{\mu_m} (2x/3)\,dx = 0.5$$

$$x^2/3 \Big|_1^{\mu_m} = 0.5 \; ; \; \mu_m = 1.58$$

ในการศึกษาการแจกจงของรายได้ในครอบครัว โดยเฉพาะในประเทศด้อยพัฒนา มัธยฐานจะอธิบายหรือแสดงคุณลักษณะของการแจกแจงรายได้ดีกว่าอย่างอื่น เพราะครอบครัวที่มีรายได้มากมีอยู่น้อย

สำหรับทางการศึกษาและจิตวิทยา เราใช้มัธยฐานกันบ่อย ๆ ดังนั้น จึงควรสนใจเรื่องนี้ให้มากหน่อย

ง. ค่าคาดหวัง (Mean or Expected Value) เป็นส่วนเฉลี่ยเลขคณิตแบบถ่วงน้ำหนัก (Weighted, Arithmetic Mean) นั่นเอง และเป็นมาตราวัดค่าเฉลี่ยที่ไว้ใจได้ที่สุด มนุษย์รู้จักและใช้กันมากตั้งแต่โบราณ Lambert Quetelet (1796-1874) นักดาราศาสตร์และสถิติชาวเบลเยี่ยมเป็นผู้พัฒนาขึ้นมา

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่ม ค่าคาดหวังของ X ซึ่งแทนด้วย $E(X)$ หรือ μ จะกำหนดไว้ดังนี้

$$E(x) = \begin{cases} \sum\limits_{\forall x} x f(x) & \text{ตัวแปรไม่ต่อเนื่อง} \\ \int\limits_{\forall x} x f(x)\,dx & \text{ตัวแปรต่อเนื่อง} \end{cases}$$

ตัวอย่าง จำนวนเงินที่นักศึกษาใช้จ่ายสำหรับค่าอาหารกลางวันมีการแจกแจงน่าจะเป็นดังนี้

x	3	5	7	10
$f(x)$	.3	.4	.2	.1

ดังนั้นค่าคาดหวังของค่าอาหารกลางวันจะเป็น

$$E(X) = 3(.3)+5(.4)+7(.2)+10(.1)$$
$$= .9+2+1.4+1 = 5.3$$

ตัวอย่าง ให้ X แทนเวลาที่ใช้อ่านหนังสือแต่ละวันของนักศึกษาคนหนึ่งมีการแจกแจงดังนี้

$$f(x) = 2(x-1) \quad , \quad 1 \le x \le 2$$

ดังนั้นค่าคาดหวังของเวลาที่ใช้อ่านหนังสือจะเท่ากับ

$$E(X) = \int_{\forall x} x f(x)\,dx = \int_1^2 x[2(x-1)]\,dx$$
$$= \int_1^2 (2x^2-2x)\,dx = 2x^3/3 - x^2 \Big|_1^2$$
$$= (16/3-4)-(2/3-1) = 1\tfrac{2}{3}$$

ค่าคาดหวังนั้นเราจะไม่แปลความหมายในทำนองว่าเกิดขึ้นบ่อยหรือน่าจะเป็นสูง แต่จะแปลว่าโดยเฉลี่ยแล้วค่าตัวแปรเชิงสุ่มจะเป็นเท่านั้นเท่านี้ซึ่งคาดหวังนั้นจะเป็นค่าที่เกิดขึ้นบ่อยหรือไม่ก็ได้ และจะเป็นค่าของตัวแปรเชิงสุ่มหรือไม่ก็ได้ จากตัวอย่างแรกเราจะเห็นว่าค่าคาดหวังเป็น 5.3 ซึ่งไม่เป็นค่าจริงของตัวแปรเชิงสุ่มเลย

ค่าคาดหวังมีบทบาทมากในทฤษฎีความน่าจะเป็นหรือสถิติ สำหรับทฤษฎีต่อไปนี้จะเป็นส่วนหนึ่งของการประยุกต์ค่าคาดหวัง

อสมการของมาร์คอฟ (Markov's Inequality) ถ้า X เป็นตัวแปรเชิงสุ่ม ที่มีค่าไม่เป็นลบ (Non-negative Values) และมีส่วนเฉลี่ยหรือค่าคาดหวังเป็น $E(X)$ แล้ว

$$P(X \ge k) \le E(X)/k$$

ในเมื่อ k เป็นจำนวนบวกใด ๆ หรือ $k > 0$

สำหรับอสมการข้างบนนี้สามารถเขียนได้เป็น

$$P[X \ge k\,E(X)] \le 1/k$$

หรือ $P[X \le k\,E(X)] \ge 1 - 1/k$

ทฤษฎีนี้จะใช้ได้เมื่อ $k \ge 1$

ตัวอย่าง ฟังก์ชั่นน่าจะเป็นกำหนดไว้ดังนี้

x	1	2	3
$f(x)$	.30	.50	.20

$$E(X) = 1(.30)+2(.50)+3(.20)$$
$$= 1.9$$

ดังนั้น $P[X \le 2\,E(X)] = P(X \le 3.8) \ge 1-1/2 = .50$

แต่จากความเป็นจริง $P(X \leq 3.8) = 1.0$

ในการคำนวณค่าคาดหวังของตัวแปรเชิงสุ่มหรือฟังก์ชั่นของตัวแปรเชิงสุ่ม ถ้าเราใช้กฎของค่าคาดหวังต่อไปนี้จะช่วยให้ง่ายต่อการคำนวณ กฎเหล่านี้ใช้ได้กับตัวแปรเชิงสุ่มทั้งสองประเภท

(1) ถ้า $g(X)$ เป็นฟังก์ชั่นของตัวแปรเชิงสุ่ม X แล้ว

$$E[g(x)] = \begin{cases} \sum_{\forall x} [g(x)] f(x) & \text{ตัวแปรเชิงสุ่มไม่ต่อเนื่อง} \\ \int_{\forall x} [g(x)] f(x) dx & \text{ตัวแปรเชิงสุ่มต่อเนื่อง} \end{cases}$$

(2) ถ้า c เป็นค่าคงที่ แล้ว $E(c) = c$

(3) ถ้า X เป็นตัวแปรเชิงสุ่มและ c เป็นค่าคงที่ แล้ว

$$E(x+c) = E(x) + c$$

(4) ถ้า X เป็นตัวแปรเชิงสุ่ม และ c เป็นค่าคงที่ แล้ว

$$E(cX) = c\,E(x)$$

(5) ถ้า X และ Y เป็นตัวแปรเชิงสุ่มสองตัวแล้ว

$$E(X+Y) = E(X) + E(Y)$$

(6) ถ้า X และ Y เป็นตัวแปรเชิงสุ่มสองตัวที่เป็นอิสระเชิงสถิติ แล้ว

$$E(XY) = E(X)\,E(Y)$$

ตัวอย่าง สมมติตัวแปรเชิงสุ่ม X มีการแจกแจงน่าจะเป็นดังนี้

x	0	1	2
$f(x)$	.25	.50	.25

$$\begin{aligned} E(X) &= 0(.25) + 1(.50) + 2(.25) \\ &= 1 \\ E(X+5) &= (0+5)(.25) + (1+5)(.50) + (2+5)(.25) \\ &= 6.0 = 1+5 \\ &= E(X) + 5 \\ E(2X) &= (2\times 0)(.25) + (2\times 1)(.50) + (2\times 2)(.25) \\ &= 0+1+1 = 2(1) = 2E(X) \end{aligned}$$

$$Y = X^3 + 3 \quad E(Y) = E(X^2+3)$$

$$= (0^2+3)(.25) + (1^2+3)\,.50 + (2^3+3)(.25)$$

$$= 4.5 = E(X^2) + 3$$

3.8.2 **มาตราวัดตำแหน่ง** (Measures of Position) เป็นพารามิเตอร์อันดับ (Order parameters) หรือค่าแบ่ง (Partition Value) ที่แบ่งการแจกแจงของตัวแปรเชิงสุ่มออกเป็นสัดส่วนต่าง ๆ มาตราวัดตำแหน่งที่เป็นพื้นฐานของมาตราอื่นจะเรียกว่าควอนไทล์หรือแฟรคไทล์ (Quantile or Fractile) ซึ่งนิยามไว้ดังนี้

นิยาม μ_p จะเรียกว่าควอนไทล์ที่ $p (0 < p < 1)$ ของตัวแปรเชิงสุ่ม X ถ้า $P(X \leq \mu_p) \geq p$ และ $P(X \geq \mu_p) \geq 1-p$

ถ้า X ตัวแปรเชิงสุ่มแบบต่อเนื่อง และ μ_p จะหาได้จากสมการ

$$P(X \leq \mu_p) = p$$

สำหรับ (1) p ที่มีค่าเป็น 0.5 เราจะเรียก μ_p ว่ามัธยฐานของ X

(2) p ที่มีค่าเป็น 0.25, 0.50, และ 0.75 เราเรียก μ_p ว่า ควอร์ไทล์ที่ 1,2 และ 3 ของตัวแปรเชิงสุ่ม X ตามลำดับ และจะแทนด้วย Q_1 , Q_2 และ Q_3

ส่วน Q_1 และ Q_3 เรามักเรียกควอร์ไทล์ต่ำและควอร์ไทล์สูง (Lower and Upper Quartile)

(3) p ที่มีค่าเป็น 0.10, 0.20,.........., 0.90 จะเรียก μ_p ว่า เดไซล์ (Decile) ที่ 1, 2,......, 9 ของตัวแปรเชิงสุ่ม X ตามลำดับ เราจะแทนด้วย $D_1, D_2, \ldots, D_9$

(4) p ที่มีค่าเป็น 0.01, 0.02......., 0.99 เราจะเรียก μ_p ว่า เปอร์-เซ็นไทล์ หรือเซ็นไทล์ (Percentile or Centile) ที่ 1, 2, ..., 99 ของตัวแปรเชิงสุ่ม X ตามลำดับ และจะแทนด้วย

ตัวอย่าง ถ้าตัวแปรเชิงสุ่ม X มีการแจกแจงดังนี้

$$f(x) = 1/\beta \; ; \; \alpha < x < \alpha + \beta$$

แล้ว ควอนไทล์ที่ p ; $p = 0.20,\ 0.50$ และ 0.75 จะหาได้ดังนี้

เมื่อ $p = 0.20$; $\int_{\alpha}^{\mu_{.2}} \frac{1}{\beta}\, dx = 0.2$

$$x/\beta \Big|_{\alpha}^{\mu_{.2}} = 0.2, \mu_{.2} = \alpha + 0.2\beta$$

เมื่อ $p = 0.50$ หรือ $p = 0.75$ เราก็ทำได้แบบเดียวกันซึ่งจะได้

$$\mu_{.5} = \alpha + 0.5\beta$$

$$\mu_{.75} = \alpha + 0.75\beta$$

ตัวอย่าง ถ้าตัวแปรเชิงสุ่ม Y มีการแจกแจงน่าจะเป็นดังนี้

y	4	5	6	7	8
$f(y)$	.1	.3	.3	.2	.1

แล้วควอนไทล์ .2 และ .50 จะหาได้ดังนี้

(1) เนื่องจาก $P(X \leq 5) = .30 > .25$ และ $P(X \geq 5) = .90 > .75$ ดังนั้น $\mu_{.2} = 5$

(2) $\mu_{.5} = 6$ เพราะ $P(X \leq 6) = .70 > .50$ และ $P(X \geq 6) = .60 > .50$

ถ้าจะหาให้เที่ยงตรงได้โดยวิธีการเทียบดังนี้

y	4	5	6	7	8
$F(y)$	.1	.4	.7	.9	1.0

(1) $\mu_{.20} = 4.5 + \dfrac{.20 - .1}{.4 - .1} = 5.0$

(2) $\mu_{.50} = 5.5 + \dfrac{.5 - .4}{.7 - .4} = 5.83$

มาตราวัดตำแหน่งนี้บางทีก็ถือว่าเป็นมาตราวัดค่าเฉลี่ยเหมือนกันในทางการศึกษาและจิตวิทยานิยมใช้มาตราวัดสรุปประเภทนี้กันมาก ดังนั้นจึงต้องให้เข้าใจแจ่มชัด

3.5.3 มาตราวัดการกระจาย (Measure of Dispersion or Concentration) เป็นปริมาณหรือพารามิเตอร์ที่แสดงถึงการกระจัดกระจาย (Spread or Scatter) หรือแสดงความจำนวนแตกต่างระหว่างค่าที่เป็นไปได้ของตัวแปรเชิงสุ่ม มาตราวัดการกระจายที่จะพูดถึงมีดังนี้

ก. พิสัย (Range) เป็นมาตราวัดการกระจายที่หยาบแต่คำนวณได้รวดเร็ว และกำหนดไว้ดังนิยามต่อไปนี้

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่ม พิสัย R จะเป็นผลต่างระหว่างค่าของตัวแปรเชิงสุ่มที่มากสุดกับน้อยสุด นั่นคือ

$$R = \max(x_i) - \min(x_i)$$

ตัวอย่าง ตัวแปรเชิงสุ่ม X มีการแจกแจงน่าจะเป็นดังนี้

x	1	2	3	4	5
$f(x)$	.1	.2	.2	.3	.2

ดังนั้น $R = \max(x_i) - \min(x_i)$

$= 5 - 1 = 4$

ข. ส่วนเบี่ยงเบนควอร์ไทล์ (Quartile Deviation, QD) เป็นครึ่งหนึ่งของระยะทางระหว่าง Q_1 และ Q_3 นั่นคือ

$$QD = \frac{1}{2}(Q_3 - Q_1)$$

บางทีเราเรียก QD นี้ว่ากึ่งพิสัยควอร์ไทล์ (Semi-Interquartile Range) เพราะ $Q_3 - Q_1$ นั้นเป็นพิสัยของควอร์ไทล์

ตัวอย่าง ให้ตัวแปรเชิงสุ่ม X มีการแจกแจงดังนี้

$$f(x) = 1/\beta \; ; \quad 2 < x < 2+\beta$$
$$Q_1 = \mu_{.25} = 2+0.25\beta$$
$$Q_3 = \mu_{.75} = 2+0.75\beta$$

ดังนั้น $QD = \frac{1}{2}\left[(2+0.75\beta)-(2+0.25\beta)\right] = 0.25\beta$

ค. ส่วนเบี่ยงเบนเฉลี่ย (Mean Deviation, MD.) เป็นค่าเฉลี่ยสมบูรณ์ (Absolute) ของส่วนเบี่ยงเบนที่ค่าของตัวแปรเชิงสุ่มห่างจากค่าเฉลี่ยของมัน P.S.Laplace (1749-1827) เป็นผู้เสนอมาตราวัดแบบนี้

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่มซึ่งมีค่าเฉลี่ย $E(X)=\mu$ แล้ว

$$MD = E|X-\mu|$$

$\sum_{\forall x} (x-\mu) f(x)$ ตัวแปรเชิงสุ่มไม่ต่อเนื่อง

$\int_{\forall x} |x-\mu| f(x)\,dx$ ตัวแปรเชิงสุ่มต่อเนื่อง

เราจะเห็นได้ว่า MD นั้นเกี่ยวข้องกับค่าสมบูรณ์ของส่วนเบี่ยงเบน ในทางคณิตศาสตร์ถือว่ายุ่งยากและพัฒนาต่อไปได้ยาก จึงจำเป็นต้องใช้วิธีอื่นที่ดีกว่า คือความแปรปรวนหรือส่วนเบี่ยงเบนมาตรฐานซึ่งจะได้กล่าวต่อไป

ตัวอย่าง กำหนดจำนวนเงินที่ต้องใช้จ่ายต่อวันของนักศึกษาจิตวิทยาซึ่งเป็นตัวแปรเชิงสุ่มและมีการแจกแจงน่าจะเป็นดังนี้

x	10	20	30
$f(x)$	0.2	0.5	0.3

$$E(X) = 10(0.2) + 20(0.5) + 30(0.3)$$
$$= 21.00$$

ดังนั้น $MD = |10-21|(0.2)+|20-21|(0.5)+|30-21|(0.3)$
$$= 2.2 + 0.5 + 2.7 = 5.4$$

ง. **ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน** (Variance and Standard Deviation) ความแปรปรวนเป็นมาตราวัดการกระจายที่มนุษย์รู้จักและใช้กันมาก K.F.Ganss (1777-1855) เป็นผู้พัฒนาจากส่วนเบี่ยงเบนเฉลี่ย โดยการเฉลี่ยกำลังสองของส่วนเบี่ยงเบนจากค่าเฉลี่ย

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่มที่มีค่าคาดหวังแล้ว ความแปรปรวนของมัน, $V(X)=\sigma^2$, กำหนดไว้ดังนี้

$$V(X) = E(X-\mu)^2 = \begin{cases} \sum_{\text{ทุก } x} (x-\mu)^2 f(x), & \text{ตัวแปรเชิงสุ่มไม่ต่อเนื่อง} \\ \int_{\text{ทุก } x} (x-\mu)^2 f(x)\,dx, & \text{ตัวแปรเชิงสุ่มต่อเนื่อง} \end{cases}$$

ตัวอย่าง ถ้าให้ X เป็นตัวแปรเชิงสุ่มที่แทนจำนวนบุตรในครอบครัวของนักจิตวิทยาการศึกษา และ X มีการแจกแจงน่าจะเป็นดังนี้

x	0	1	2	3	4
$f(x)$	.01	.25	.50	.15	.09

ดังนั้น

$$E(X) = 0(.01)+1(.25)+2(.50)+3(.15)+4(.09)$$
$$= 0+.25+1.00+.45+.36 = 2.06$$
$$V(X) = (0-2.06)^2(.01)+(1-2.06)^2(.25)+(2-2.06)^2(.50)$$
$$+(3-2.06)^2(.15)+(4-2.06)^2(.09)$$
$$= 7964$$

ในการคำนวณความแปรปรวนจากตัวอย่างข้างบนจะยุ่งยาก เราจะดัดแปลงนิยามข้างบนซึ่งจะได้สูตรต่อไปนี้

$$V(X) = E(X-\mu)^2$$
$$= E(X^2)-\mu^2$$

สูตรนี้จะสะดวกต่อการคำนวณเป็นอย่างมาก จากตัวอย่างข้างบนนี้ ถ้าเราใช้สูตรใหม่นี้ก็จะได้

$$V(X) = \left[0^2(.01)+1^2(.25)+2^2(.50)+3^2(.15)\right.$$
$$\left.+4(.09)\right]-(2.06)^2$$
$$= \left[0+.25+2.00+1.35+1.44\right]-4.2436$$
$$= 5.04-4.2436 = 0.7964$$

คุณสมบัติของความแปรปรวน (Laws of Variance) ถ้ากำหนดให้ c เป็นค่าคงที่ และ X เป็นตัวแปรเชิงสุ่มใด ๆ ที่ $V(X) \geqslant 0$ แล้ว เราจะได้กฎต่าง ๆ ต่อไปนี้ ซึ่งจะเป็นประโยชน์ในการคำนวณความแปรปรวน

(1) ถ้า $g(X)$ เป็นฟังก์ชั่นของ X แล้ว $V[g(X)]$ กำหนดไว้ดังนี้

$$V[g(X)] = E\{g(X)-E[g(X)]\}^2$$
$$= E[g(X)]^2-\{E[g(X)]\}^2$$

(2) $V(C) = 0$

(3) $V(X+C) = V(X)$

(4) $V(CX) = C^2V(X)$

(5) ถ้า X และ Y เป็นตัวแปรเชิงสุ่มที่เป็นอิสระกันแล้ว

$$V(X \pm Y) = V(X)+V(Y)$$

(6) $V(X) \leq E(X-c)^2$

ความแปรปรวนเป็นมาตราวัดการกระจายที่เชื่อถือได้เช่นเดียวกับค่าคาดหวังที่เป็นมาตราวัดค่าเฉลี่ย ในทางธุรกิจใช้ความแปรปรวนเป็นมาตราวัดการเสี่ยง (Risk) ในทางจิตวิทยาและการศึกษาจะใช้เป็นมาตราวัดคุณภาพของเครื่องมือ วิธีการหรือกรรมวิธีทดลองได้

ในการใช้ความแปรปรวนจะมีความยุ่งยากประการหนึ่งคือหน่วยของมันจะเป็นกำลังสองของมิติของตัวแปรเชิงสุ่ม เช่นถ้า X เป็นนิ้ว $E(X)$ จะเป็นนิ้ว แต่ $V(X)$ จะเป็นนิ้วกำลังสอง เพื่อให้มาตราวัดนี้มีหน่วยเดียวกับตัวแปรเชิงสุ่ม เราจึงถอดรากที่สองของความแปรปรวนเสีย และจะเรียกมันใหม่ว่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรเชิงสุ่ม

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่มที่มีความแปรปรวน $V(X)$ แล้วส่วนเบี่ยงเบนมาตรฐาน σ จะกำหนดไว้ดังนี้

$$\sigma = \sqrt{V(X)}$$

$$= \begin{cases} \sqrt{E(X-\mu)^2} \\ \sqrt{E(X^2) - \mu^2} \end{cases}$$

ตัวอย่าง ถ้าให้ X เป็นตัวแปรเชิงสุ่มที่แทนเวลาที่นักศึกษาจิตวิทยาแต่ละคนใช้สอบวิชาสถิติจิตวิทยาและ X มีการแจกแจงดังนี้

$$f(x) = 2x/5 \ , \ 2 \leq x \leq 3$$

เราจะได้ $$E(X) = \int_2^3 x(2x/5)\,dx = 2x^3/15 \Big|_2^3 = 38/15$$

$$E(X^2) = \int_2^3 x^2(2x/5)\,dx = 2x^4/20 \Big|_2^3 = 6.5$$

ดังนั้น $$V(X) = 6.5 - (38/15)^2 \cong 0.10$$

$$\sigma = \sqrt{0.10} \cong 0.3$$

บทบาทของส่วนเบี่ยงเบนมาตรฐานในฐานะมาตราวัดการกระจายมีความสำคัญมากในเมื่อ P.L.Tchebycheff (1881-1894) นักคณิตศาสตร์ชาวรัสเซียได้พัฒนาทฤษฎีโดยอาศัยมาตราวัดนี้ และได้ชื่อว่า อสมการเชบบีเชฟ (Tchebycheff's Ineguality) ซึ่งกล่าวไว้ดังนี้

ทฤษฎี ความน่าจะเป็นที่ค่าของตัวแปรเชิงสุ่ม X จะเบี่ยงเบนจากค่าคาดหวังของมันมากกว่า k ส่วนเบี่ยงเบนมาตรฐาน $(k\sigma)$ นั้นจะเท่ากับหรือน้อยกว่าส่วนกลับของกำลังสอง k นั่นคือ

$$P(|X - E(X)| > k\sigma) \leq 1/k^2$$

จากทฤษฎีนี้เราจะได้อีกว่า

$$P(|X - E(X)| \leq k\sigma \geq 1 - 1/k^2$$

ถ้า k อยู่ระหว่าง 0 กับ 1 จะไม่ให้อะไรเพราะจะได้ความน่าจะเป็นที่เป็นลบซึ่งเป็นไปไม่ได้ ลองพิจารณาความน่าจะเป็นจากอสมการข้างบนนี้โดยที่ k มีค่าต่าง ๆ กัน

k	ช่วง	ความน่าจะเป็น
1.25	$\mu \pm 1.25\sigma$	0.36
1.50	$\mu \pm 1.50\sigma$	0.56
2.00	$\mu \pm 2.00\sigma$	0.75
2.50	$\mu \pm 2.50\sigma$	0.84
3.00	$\mu \pm 3.00\sigma$	0.89
4.00	$\mu \pm 4.00\sigma$	0.94

ตัวอย่าง ถ้าตัวแปรเชิงสุ่ม X มีการแจกแจงน่าจะเป็นดังนี้

X	1	2	3	4
$f(x)$	.2	.3	.3	.2

แล้ว

$$E(X) = (1-2.5)^2(.2) + (2-2.5)^2(.3) + (3-2.5)^2(.3) + (4-2.5)^2(.2)$$
$$= 1.05$$

และ $\sigma = \sqrt{1.05} = 1.02$

ถ้าเราสนใจจะหาความน่าจะเป็นที่ค่าของ X จะอยู่ห่างจากค่าคาดหวังไม่เกิน 1.25 เราก็หาได้จากทฤษฎี นั่นคือ

$$P\{|X - 2.5| \leq 1.25(1.02)\} \geq 0.36$$

แต่จากที่เป็นจริงเราจะได้ว่า

$$P\{|X - 2.5| \leq 1.25(1.02)\}$$
$$= P \quad 1.225 \leq X \leq 3.775$$
$$= 0.6$$

ซึ่งก็เป็นจริงตามทฤษฎี

ในทางปฏิบัติหรือการทดลองเราได้กฎการทดลองที่สรุปผลไว้ดังนี้

กฎการทดลอง (Empirical Rule) ถ้าตัวแปรเชิงสุ่มมีการแจกแจงแบบระฆัง (Bellshaped) แล้วโอกาสที่ค่าของตัวแปรเชิงสุ่มจะอยู่ในช่วง $\mu \pm \sigma$, และ $\mu \pm 2\sigma$ จะประมาณ 68, 95, และ 99.7% ตามลำดับ

จ. สัมประสิทธิ์การกระจาย (Coefficient of Dispersion) มาตราวัดการกระจายแบบสัมพัทธ์ซึ่งจะเป็นเลขจำนวนแท้ที่ไม่มีหน่วย และกำหนดไว้ดังนี้

$$\text{สัมประสิทธิ์การกระจาย} = \frac{\text{มาตราวัดการกระจาย}}{\text{มาตราวัดค่าเฉลี่ยที่เหมาะสม}}$$

โดยที่สัมประสิทธิ์การกระจายเป็นมาตราวัดที่ไม่มีหน่วย เราจึงใช้เปรียบเทียบการกระจายของสองตัวแปรเชิงสุ่มได้ถึงแม้ว่าตัวแปรเชิงสุ่มทั้งสองจะมีขนาดของค่าที่เป็นไปได้แตกต่างกัน หรือมีหน่วยของค่าตัวแปรเชิงสุ่มต่างกันได้ เช่นค่าของตัวแปรเชิงสุ่มตัวหนึ่งมีหน่วยเป็นพันบาท อีกตัวหนึ่งมีหน่วยเป็นสิบบาท หรือตัวหนึ่งมีหน่วยเป็นดวลล่าร์ แต่อีกตวหนึ่งมีหน่วยเป็นรูปี เป็นต้น สัมประสิทธิ์การกระจายที่สำคัญมีดังนี้

(1) สัมประสิทธิ์ความผันแปร (Coefficient of Vaviation, CV) มาตราวัดนี้ Karl Pearson (1857-1936) เป็นผู้พัฒนาขึ้นมาดังสูตรต่อไปนี้

$$CV = \frac{\sigma}{\mu}(100)$$

ถ้า $\mu \longrightarrow 0$ แล้ว CV จะเป็นมาตราวัดที่ไม่ดี

ตัวอย่าง สมมติว่าตัวแปรเชิงสุ่ม X และ Y มีค่าคาดหวังและส่วนเบี่ยงเบนมาตรฐานดังนี้

	ค่าคาดหวัง	ส่วนเบี่ยงเบนมาตรฐาน
X	2,000,000	200,000
Y	1,000,000	10,000

ตัวแปรเชิงสุ่ม X, $CV = \frac{2,000,000}{200,000}(100) \quad 10\%$

ตัวแปรเชิงสุ่ม Y, $CV = \frac{100,000}{10,000}(100) \quad 10\%$

ถ้ามองเผิน ๆ จะเห็นว่าตัวแปรเชิงสุ่ม X จะมีการกระจายมากกว่า แต่ที่จริงทั้งสองมีการกระจายเท่ากันดังที่หาได้

(2) สัมประสิทธิ์ส่วนเบี่ยงเบนควอร์ไทล์ (Coefficient of Quartile Deviation, V_Q) มาตราวัดการกระจายนี้กำหนดไว้ดังนี้

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}(100)$$

(3) สัมประสิทธิ์ส่วนเบี่ยงเบนเฉลี่ย (Coefficient of Mean Deviation, Vm) กำหนดไว้ดังนี้

$$V_m = \begin{cases} \frac{MD}{\mu}(100) \\ \frac{MD}{\mu_m}(100) \end{cases}$$

3.6 โมเมนต์ของตัวแปรเชิงสุ่ม (Moment)

โมเมนต์ของตัวแปรเชิงสุ่มก็คือค่าคาดหวังของกำลังของตัวแปรเชิงสุ่มนั่นเอง และมันจะอธิบายคุณลักษณะของฟังก์ชั่นน่าจะเป็นได้ดีเพราะโมเมนต์ของฟังก์ชั่นน่าจะเป็นหนึ่ง ๆ จะมีเพียงตัวเดียว (Unique)

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่มแล้วค่าคาดหวังของฟังก์ชั่น $q(x) = X^k$ หรือ $m_k = E(X^k);\ k=0,1,\ldots$ จะเรียกว่าโมเมนต์ที่ หรือโมเมนต์รอบจุดกำเนิดของตัวแปรเชิงสุ่ม X

ค่าคาดหวังและความแปรปรวนที่กล่าวมาแล้วก็เป็นโมเมนต์อย่างหนึ่งนั่นคือเมื่อ

$$k = 1,\quad m_1 = E(X) = \mu$$
$$k = 2\quad m_2 = E(X^2)$$

ดังนั้น $$V(x) = E(X^2) - \mu^2 = m_2 - m_1^2$$

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่ม แล้ว m_k^1 จะเรียกว่าโมเมนต์ที่ k รอบจุด C ใด ๆ ถ้ากำหนดไว้ดังนี้

$$m_k^1 = E(X-C)^k;\quad k = 1, 2, 3, \ldots$$

ถ้า $C = \mu$ หรือ m_1 จะเรียกว่าโมเมนต์กลาง (Central moment) และจะแทนด้วย μ_k ดังนี้

$$\mu_k = E(X-\mu)^k\ ;\quad k = 1, 2, 3 \ldots$$

เมื่อ
$$k = 1,\quad \mu_1 = 0$$
$$k = 2\quad \mu_2 = V(X)$$
$$k = 3\quad \mu_3 =$$
ใช้เป็นมาตราวัดความไม่สมมาตราของตัวแปรเชิงสุ่ม

$k = 4,\ \mu_4$ = ความสูง (Krutosis) ใช้เป็นมาตราวัดความสูง (Peaked or flat) ของการแจกแจงของตัวแปรเชิงสุ่ม

นิยาม ถ้า X เป็นตัวแปรเชิงสุ่ม แล้ว μ_k' จเรียกว่าโมเมนต์ผลคูณ (Factarial moment) ของ X ถ้ากำหนดไว้ดังนี้

$$\mu'_k = E[X(k)]$$
$$= E[X(X-1)\ldots\ldots(X-k+1)]$$

เมื่อ

$$k = 1 \;,\; \mu'_1 = E(X)$$
$$k = 2 \;,\; \mu'_2 = E[X(X-1)] = E(X^2) - E(X)$$
$$= m_2 - m_1$$

นิยาม โมเมนต์สมบูรณ์ (Absolute moment) รอบจุด จุดกำเนิด และรอบค่าคาดหวัง ของตัวแปรเชิงสุ่ม X จะกำหนดไว้ดังนี้

$$V_k = E|X|^k \;;\quad k = 1,2,3,\ldots\ldots$$

และ $V_k = E|X-\mu|^k \;;\; k = 1,2,3,\ldots\ldots$

เมื่อ $k = 1 \;,\; V_1 = E|X-\mu| = MD.$

ตัวอย่าง สมมติตัวแปรเชิงสุ่ม X มีการแจกแจงน่าจะเป็นดังนี้

x	1	2	3	4
$f(x)$	.2	.3	.3	.2

แล้วโมเมนต์ต่าง ๆ จะหาได้ดังนี้ (1)

(1) $m_k = E(X^k)$

$$= 1^k(.2) + 2^k(.3) + 3^k(.3) + 4^k(.2)$$
$$= (.2)[1+4^k] + (.3)[2^k+3^k] \;;\; k=0,1,2,\ldots\ldots$$

(2) $m'_k = E(X-C)^k$

$$= (1-C)^k(.2) + (2-C)^h(.3) + (3-C)^k(.3) + (4-C)^k(.2)$$
$$= [(1-C)^k + (4-C)^k](.2) + [(2-C)^k + (3-C)^k](.3)$$
$$k = 1,2,\ldots\ldots\ldots$$

(3) $E(X) = \mu = 2.5$

$$\mu_k = E(X-\mu)^k$$
$$= (1-2.5)^k(.2) + (2-2.5)^k(.3) + (3-2.5)^k(.3) + (4-2.5)^k(.2)$$
$$= (.2)[(-1.5)^k + (1.5)^k] + (.3)[(-.5)^k + (.5)^k] \;;$$
$$k = 1,2,3\ldots\ldots$$

(4) $\mu'_k = E[X^{(k)}]$

สำหรับ $k = 2,\quad \mu' = E[X(X-1)]$

$$= (1)(1-1)(.2) + (2)(2-1)(.3) + (3)(3-1)(.3) + (4)(4-1)(.2)$$

$$= 0 + .6 + 1.8 + 2.4 \quad = 4.8$$

(5) $$V_k' = E|X|^k$$
$$= |1|^k(.2) + |2|^k(.3) + |3|^k(.3) + |4|^k(.2)$$
$$= (.2)\left[1 + 4^k\right] + (.3)\left[2^k + 3^k\right]; \; k = 1, 2, 3, \ldots\ldots$$

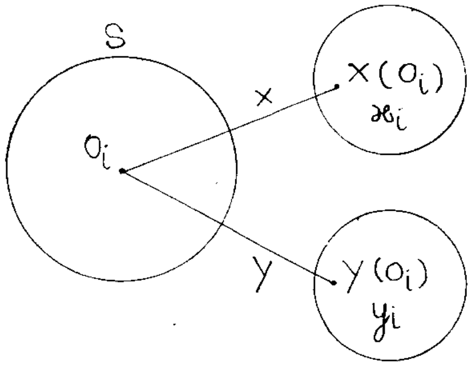
(6) $$V_k = E|X - \mu|^k$$
$$= |1 - 2.5|^k(.2) + |2 - 2.5|^k(.3) + |3 - 2.5|^k(.3)$$
$$+ |4 - 2.5|^k(.2)$$
$$= (.2)(1.5)^k(2) + (.3)(.5)^k(2)$$
$$= (2)\left[(.2)(1.5)^k + (.3)(.5)^k\right]; \; k = 1, 2, 3, \ldots\ldots$$

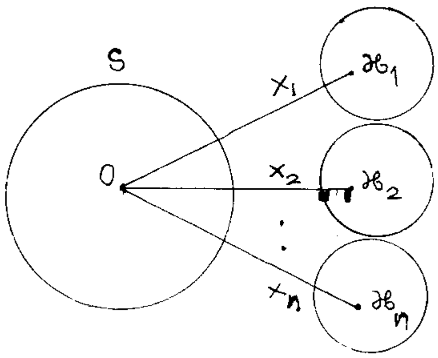
3.7 การแจกแจงหลายตัวแปร (Multivariate Distributions)

ในการพิจารณาผลทดลองเชิงสุ่มเท่าที่ผ่านมาเราพิจารณาเพียงคุณลักษณะเดียวเท่านั้น ตัวแปรเชิงสุ่มที่กำหนดในกลุ่มผลทดลองเช่นนั้นเราเรียกว่าตัวแปรแล้วสมมติเดียว (One-dimensional nandam variables) บ่อยครั้งที่เราสนใจผลทดลองมากกว่าหนึ่งคุณลักษณะ เช่นการวิเคราะห์ผู้สมัครนั้นจำเป็นต้องพิจารณาระดับการศึกษา (E) และระดับสติปัญญา (I) ซึ่งเราจะได้ผลทดลอง (e, i) เป็นต้น ตัวแปรเชิงสุ่มที่กำหนดในกลุ่มผลทดลองอันประกอบด้วยผลทดลอง 2,3,.....,**n** คุณลักษณะนั้นเราเรียกว่า **ตัวแปรเชิงสุ่ม** 2, 3,...,**n นิติ** ลองศึกษานิยามต่อไปนี้

นิยาม ให้ S เป็นกลุ่มผลทดลองของการทดลองเชิงสุ่ม ถ้า x และ y เป็นฟังก์ชั่นที่กำหนดเลขจำนวนจริง $X(O_i)$ และ $Y(O_i)$ หรือ(x_i, y_i) ให้แต่ละผลทดลอง $O_i \in S$ เราจะเรียก (X, Y) ว่าตัวแปรเชิงสุ่มสองมิติหรือเวคเตอร์เชิงสุ่มสองมิติ (Two-dimensional Random Variable or Vector)

ถ้า $X_1, X_2, \ldots\ldots, X_n$ เป็น n ฟังก์ชั่นที่กำหนดเลขจำนวนจริง$(x_1, x_2, \ldots, x_n)$ หรือ $\underline{X}$ ให้แก่แต่ละผลทดลอง $O_i \in S$ เราจะเรียก $(X_1, X_2, \ldots, X_n)$ หรือ $\underline{X}$ ว่าตัวแปร (เวคเตอร์) เชิงสุ่ม n มิติ





ตัวแปรเชิงสุ่มสายตัวนี้เราแบ่งได้เป็น 2 ประเภท คือ แบบต่อเนื่องและไม่ต่อเนื่อง วึ่งเป็นเช่นเดียวกับกรณีของตัวแปรเชิงสุ่มตัวเดียว

นิยาม ก. (X, Y) จะเป็นตัวแปรเชิงสุ่มแบบไม่ต่อเนื่องสองมิติ ถ้าค่าที่เป็นไปได้ของ (X, Y) มีจำกัดหรืออนันต์นับได้ นั่นคือค่าที่เป็นไปได้ของ (X, Y) จะแทนได้เป็น (x_i, y_i) ในทำนองเดียวกัน $(X_1, X_2, \ldots, X_n$ หรือ $\underline{X}$ จะเป็นตัวแปรเชิงสุ่ม n มิติแบบไม่ต่อเนื่อง ถ้าค่าที่เป็นไปได้ของ $\underline{X}$ มีจำกัดหรือเป็นอนันต์นับได้

ข. (X, Y) จะเป็นตัวแปรเชิงสุ่มต่อเนื่อง 2 มิติ ถ้า (X, Y) สามารถมีได้ทุกค่าในเซ็ทที่นับไม่ได้ของระนาบ (Non-countable Set of the Euclidean Plane)

เช่นเดียวกัน $\underline{X}$ จะเป็นตัวแปรเชิงสุ่มต่อเนื่อง n มิติ ถ้าค่าที่เป็นไปได้ของ $\underline{X}$ สามารถมีได้ทุก ๆ ค่าในเขตของกลุ่มnมิติ (Region of the n-dimensional space)

ในกรณีที่เราเกี่ยวข้องกับตัวแปรเชิงสุ่ม n มิติ นอกจากจะสนใจคุณลักษณะแต่ละอย่างแยกกัน เรายังจะสนใจถึงความสัมพันธ์ภายใน (Interrelationships) ที่มีอยู่ระหว่างคุณลักษณะต่าง ๆ นั้น และการที่เราพิจารณาถึงความสัมพันธ์ระหว่างคุณลักษณะต่าง ๆ นั้นจำเป็นที่จะต้องศึกษาหรือสืบสวนฟังก์ชั่นน่าจะเป็นของการเกิดร่วมกันหรือการแจกแจงร่วมของมันด้วย ซึ่งเราจะให้นิยามต่อไปนี้ (เพื่อความง่ายต่อการศึกษาเราจะสนใจตัวแปรเชิงสุ่ม 2 มิติเท่านั้น)

นิยาม ก. ให้ (X, Y) เป็นตัวแปรเชิงสุ่ม 2 มิติแบบไม่ต่อเนื่อง ฟังก์ชั่น f ที่กำหนดความน่าจะเป็น $f(x, y) = P(X = x), Y = y$ ให้แก่ค่า (x, y) ของตัวแปรเชิงสุ่ม (X, Y) นั้นเราจะเรียกว่าการแจกแจงน่าจะเป็นร่วม (joint probpbility distribution) ของ (X, Y) ถ้า $f(x, y)$ สอดคล้องกับเงื่อนไขต่อไปนี้

(1) $f(x, y) \geq 0 \quad \forall (x, y)$

(2) $\sum\limits_{x, y} f(x, y) = 1$

ข. ให้ (X, Y) เป็นตัวแปรเชิงสุ่ม 2 มิติแบบต่อเนื่องและมีค่าที่เป็นไปได้อยู่ในเขตของระนาบ R ฟังก์ชั่น f ที่กำหนดจำนวนเลข (x, y) ให้แก่ค่า (x, y) ของตัวแปรเชิงสุ่ม (X, Y) นั้น เราจะเรียกว่า การแจกแจงหนาแน่นน่าจะเป็นร่วม (Joint prebability density function) ของ (X, Y) ถ้า $f(x, y)$ สอดคล้องกับเงื่อนไขต่อไปนี้

(1) $f(x, y) \geq 0 \quad \forall (x, y) \in R$

(2) $\int_R \int f(x, y)\, dx dy = 1$

สำหรับตัวแปรเชิงสุ่ม 2 มิติแบบต่อเนื่องนี้ เราจะกำหนดความน่าจะเป็นได้ดังนี้

$P(x \leq X \leq x + \Delta x, \; y \leq Y \leq y + \Delta y) = f(x, y) \Delta x \Delta y$, ถ้า Δx และ Δy เล็กพอ

หรือ $P(a \le x \le b,\ c \le y \le d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$; ถ้า $a < b$ และ $c < d$

ตัวอย่าง ชมรมจิตวิทยาต้องการเลือกตัวแทน 3 คน โดยการเลือกสุ่มจากคณะกรรมการบริหารชมรม 10 คน เพื่อไปประชุมสัมมนาทางด้านจิตวิทยากับมหาวิทยาลัยอื่น ๆ คนแรกที่สุ่มมาได้จะทำหน้าที่หัวหน้าคณะ คณะกรรมการบริหารชมรมนี้ครึ่งหนึ่งเป็นนักศึกษาหญิง

ถ้าให้ M และ F แทนนักศึกษาชายและหญิงตามลำดับ แล้วเราจะได้กลุ่มผลทดลองของการเลือกตัวแทน 3 คน ดังนี้

$$S = \{MMM, MMF, MFM, FMM, MFF, FMF, FFM, FFF\}$$

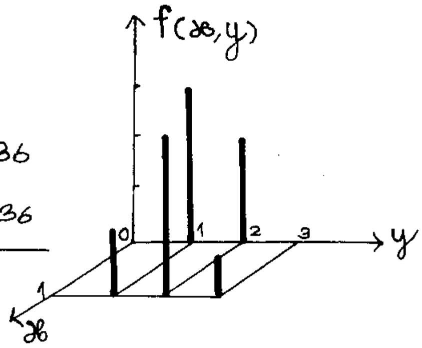
และถ้าให้ x และ y เป็นฟังก์ชั่นที่กำหนดในกลุ่มผลทดลองนี้โดยให้ x แทนหัวหน้าคณะซึ่งจะมีค่าเป็น 0 ถ้าเป็นชาย และเป็น 1 เมื่อเป็นหญิง และ y แทนจำนวนนักศึกษาหญิงในผู้แทน 3 คนนั้น ซึ่งจะมีค่าเป็น 0, 1, 2, 3

เราสามารถแสดงผลทดลอง, ค่าของ x และ y และความน่าจะเป็นที่สมนัยกับผลทดลองหรือค่าของ x และ y ได้ดังนี้

ผลทดลอง (O_i)	x	y	**ความน่าจะเป็น,** $P(O_i)$
MMM	0	0	$\left(\frac{5}{10}\right)\left(\frac{4}{9}\right)\left(\frac{3}{8}\right) = 3/36$
MMF	0	1	$\left(\frac{5}{10}\right)\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) = 5/36$
MFM	0	1	$\left(\frac{5}{10}\right)\left(\frac{5}{9}\right)\left(\frac{4}{8}\right) = 5/36$
.	.		
.	.		
.	.		
FFM	1	2	$\left(\frac{5}{10}\right)\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) = 5/36$
FFF	1	3	$\left(\frac{5}{10}\right)\left(\frac{4}{9}\right)\left(\frac{3}{8}\right) = 3/36$

จากตารางข้างบนเราสามารถสร้างฟังก์ชั่นน่าจะเป็นร่วมของ (x, y) โดยตารางและกราฟได้ดังนี้

$x \backslash y$	0	1	2	3	
0	3/36	10/36	5/36	0	18/36
1	0	5/36	10/36	3/36	18/36
	3/36	15/36	15/36	3/36	1



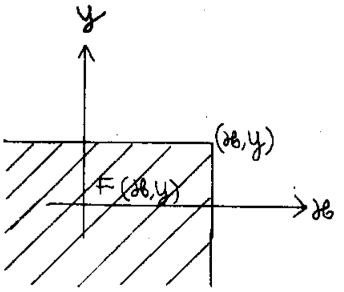
3.8 ฟังก์ชั่นแจกแจงสะสมของตัวแปรเชิงสุ่มสองมิติ (Cumulative Distribution Functions, CDF)

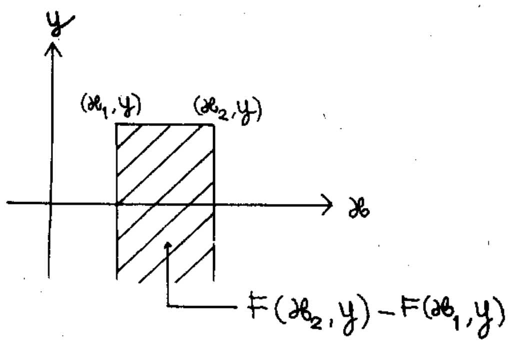
นิยาม ให้ (X,Y) เป็นตัวแปรเชิงสุ่ม 2 มิติที่มีค่าเป็น (x,y) ฟังก์ชั่น F ที่กำหนดความน่าจะเป็น, $F(x,y)$,ให้แก่เหตุการณ์ที่ X มีค่าน้อยกว่าหรือเท่ากับ x และ Y มีค่าน้อยกว่าหรือเท่ากับ y นั้นจะเรียกว่าฟังก์ชั่นแจกแจงสะสมของตัวแปรเชิงสุ่ม (X,Y) นั่นคือ

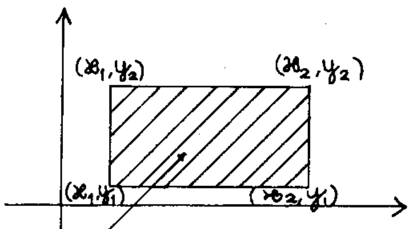
$$F(x,y) = P(X \le x,\ Y \le y)$$

$$= \begin{cases} \sum\limits_{X \le x} \sum\limits_{Y \le y} f(x,y) \\ \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{y} f(x,y)\, dx\, dy \end{cases}$$

ฟังก์ชั่นแจกแจงสะสมร่วมของตัวแปรเชิงสุ่ม 2 มิติ และความน่าจะเป็นในเทอมของฟังก์ชั่นแจกแจงสะสมร่วม แสดงได้ดังนี้







$$P(x_1 < X < x_2,\ y_1 < Y < y_2) = [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)]$$

ถ้า F เป็นฟังก์ชั่นแจกแจงสะสมของตัวแปรเชิงสุ่ม 2 มิติ (X,Y) ที่มีฟังก์ชั่นหนาแน่นน่าจะเป็นร่วม f แล้ว

$$\frac{\partial^2 F(x,y)}{\partial x\, \partial y} = f(x,y);$$ ในเมื่อ F หาอนุพันธ์ได้

ตัวอย่าง กำหนดฟังก์ชั่น $$f(x,y) = (3-x-y)/3 \quad ; \quad \begin{matrix} 0 \le x \le 2 \\ 0 \le y \le 1 \end{matrix}$$

เราสามารถหา $F(x,y)$ ได้ดังนี้

$$F(x,y) = \int_0^x \int_0^y [(3-x-y/3]\,dx\,dy$$

$$= \frac{1}{3}\left(3xy - \frac{xy^2}{2} - \frac{x^2y}{2}\right)$$

ดังนั้น $P(X \le 2, Y \le 1) = F(2,1) = 1$

และ $P(X < 1, Y < 1) = F(1,1) = 2/3$

3.9 ฟังก์ชั่นน่าจะเป็นทางเดียวและแบบเงื่อนไข (Marginal and Conditional Probability functions)

ถึงแม้ว่าเราจะมีฟังก์ชั่นน่าจะเป็นร่วมสำหรับ (X,Y) เราก็ยังสนใจที่จะหาฟังก์ชั่นของตัวแปรเชิงสุ่มแต่ละตัวแยกกัน และฟังก์ชั่นของแต่ละตัวจากตัวแปรเชิงสุ่มที่มีการแจกแจงร่วม เราจะเรียกว่าฟังก์ชั่นน่าจะเป็นทางเดียว (Marginal probability function) ซึ่งจะนะยามต่อไปนี้

นิยาม ถ้า f เป็นฟังก์ชั่นน่าจะเป็นร่วมของตัวแปรเชิงสุ่ม (X,Y) แล้ว g และ h จะเป็นฟังก์ชั่นน่าจะเป็นทางเดียวของ X และ Y ถ้ากำหนดไว้ดังนี้

$$f(x) = \begin{cases} \sum_y f(x,y) \\ \int_y f(x,y)\,dx\,dy \end{cases}$$

และ $$f(y) = \begin{cases} \sum_x f(x,y) \\ \int_x f(x,y)\,dx\,dy \end{cases}$$

ตัวอย่าง ถ้า X และ Y เป็นตัวแปรเชิงสุ่มที่มีฟังก์ชั่นร่วมดังนี้

$$f(x,y) = 1/4 \;;\; (x,y) = (-1,-1), (-1,1), (1,-1), (1,1)$$

แล้ว $$f(x) \quad \begin{matrix} f(-1,-1) + f(-1,1) = \frac{1}{2} \;; & x = -1 \\ f(1,-1) + f(1,1) = \frac{1}{2} \;; & x = 1 \end{matrix}$$

หรือ $f(x) = 1/2 \;;\; x = -1, 1$

ในทำนองเดียวกัน $f(y) = 1/2, \; y = -1, 1$

ตัวอย่าง ถ้าฟังก์ชั่นร่วมของ (X,Y) กำหนดไว้ดังนี้

$$f(x,y) = (6-x-y)/8 \quad ; \quad \begin{array}{l} 0 \le x \le 2 \\ 2 \le y \le 4 \end{array}$$

แล้ว $f(x) = \int_2^4 \frac{1}{8}(6-x-y)\,dy = \frac{1}{4}(3-x)\,;\ 0 \le x \le 2$

และ $f(y) = \int_0^2 \frac{1}{8}(6-x-y)\,dx = \frac{1}{4}(5-y)\,;\ 2 \le y \le 4$

สำหรับตัวแปรเชิงสุ่ม **(X,Y)** นั้นเมื่อเราพิจารณาฟังชันน่าจะเป็นทางเดียวของ X เราก็พิจารณาเฉพาะค่าฟังก์ชั่นของ X โดยไม่สนใจหรือเกี่ยวข้องกับค่าของ Y แต่ถ้าเราสนใจเฉพาะเจาะจงค่าของ Y ไว้เช่น $Y=y$ ฟังก์ชั่นน่าจะเป็นของ X เมื่อกำหนด $Y=y$ นี้จะเรียกว่า **ฟังก์ชั่นน่าจะเป็นเงื่อนไขของ X เมื่อกำหนด $Y=y$ ให้** ดังนิยามต่อไปนี้

นิยาม ให้ (X,Y) เป็นตัวแปรเชิงสุ่ม 2 มิติ ที่มีฟังก์ชั่นน่าจะเป็นร่วม f และให้ g, h เป็นฟังก์ชั่นน่าจะเป็นทางเดียวของ X และ Y ตามลำดับ แล้วฟังชั่นก็น่าจะเป็นเงื่อนไขของ X กำหนด $Y=y$ หรือ $f(x/y)$ กำหนดไว้ดังนี้

$$f(x/y) = \frac{f(x,y)}{f(y)} \quad ; \quad f(y) > 0$$

และฟังก์ชั่นน่าจะเป็นเงื่อนไขของ Y กำหนด $X=x, f(y|x)$ ก็กำหนดไว้ดังนี้

$$f(y/x) = \frac{f(x,y)}{f(x)} \quad ; \quad f(x) > 0$$

จากนิยามข้างบนนี้เราจะได้กฎของการคูณเกี่ยวกับฟังก์ชั่นดังนี้

กฎการคูณ $f(x,y) = \begin{cases} f(x/y)\,f(y) \\ f(y/x)\,f(x) \end{cases}$

ตัวอย่าง ถ้า (X,Y) มีการแจกแจงร่วมดังนี้

$X \backslash Y$	y_1	y_2	y_3	$f(x_i)$
x_1	.05	.10	.20	.35
x_2	.20	.20	.05	.45
x_3	.05	.05	.10	.20
$f(y_j)$	.30	.35	.35	1.00

แล้วฟังก์ชั่นน่าจะเป็นเงื่อนไขของ x กำหนด $y = y$ จะเป็นดังนี้

$$f(x/y) = f(x,y) / f(y)$$

$$f(x/y_1) = \frac{.05}{.30} = 1/6 \; ; \; x = x_1$$

$$= \frac{.20}{.30} = 4/6 \; ; \; x = x_2$$

$$= \frac{.05}{.30} = 1/6 \; ; \; x = x_3$$

$$f(x/y_2) = .10/.35 = 2/7 \; ; \; x = x_1$$

$$= .20/.35 = 4/7 \; ; \; x = x_2$$

$$= .05/.35 = 1/7 \; ; \; x = x_3$$

$$f(x/y_3) = .20/.35 = 4/7 \; ; \; x = x_1$$

$$= .05/.35 = 1/7 \; ; \; x = x_2$$

$$= .10/.35 = 2/7 \; ; \; x = x_3$$

เราสรุปได้ดังตาราง

x	x_1	x_2	x_3
$f(x\|y_1)$	1/6	4/6	1/6
$f(x\|y_2)$	2/7	4/7	1/7
$f(x\|y_3)$	4/7	1/7	2/7

และฟังก์ชั่นน่าจะเป็นเงื่อนไขของ y กำหนด $x = x$ จะเป็นดังนี้

$$f(y/x) = f(x,y)/f(x)$$

ซึ่งจะทำได้เช่นเดียวกับข้างบน และผลสุดท้ายจะเป็นดังตารางต่อไปนี้

y	y_1	y_2	y_3
$f(y/x_1)$	1/7	2/7	4/7
$f(y/x_2)$	4/9	4/9	1/9
$f(y/x_3)$	1/4	1/4	2/4

ตัวอย่าง จากตัวอย่างที่ผ่านมาเรามี

$$f(x,y) = \frac{1}{8}(6-x-y) \quad ; \quad \begin{matrix} 0 \le x \le 2 \\ 2 \le y \le 4 \end{matrix}$$

$$f(x) = \frac{1}{4}(3-x)$$

และ

$$f(y) = \frac{1}{4}(5-y)$$

แล้วเราจะได้

$$f(x/y) = f(x,y)/f(y) = (6-x-y)/2(5-y)$$

และ

$$f(y/x) = f(x,y)/f(x) = (6-x-y)/2(3-x)$$

3.10 ความเป็นอิสระเชิงสถิติของตัวแปรเชิงสุ่ม (Statistical Independence)

จากความน่าจะเป็นเบื้องต้นเราได้นิยามความเป็นอิสระระหว่างสองเหตุการณ์ A และ B ไว้แล้ว แนวความคิดในเรื่องความเป็นอิสระของตัวแปรเชิงสุ่มสองตัว (X,Y) ก็เช่นเดียวกัน คือเราจะพูดว่า X เป็นอิสระจาก Y ถ้าผลทดลองของ X ไม่มีอิทธิพลต่อผลทดลองของ Y ดังนิยาม

นิยาม ตัวแปรเชิงสุ่ม 2 มิติ (X,Y) จะเป็นอิสระกันก็ต่อเมื่อ

$$f(x/y) = f(x)$$

หรือ

$$f(y/x) = f(y)$$

ดังนั้นจากกฎการคูณของฟังก์ชั่นเราจะได้ทฤษฎีเกี่ยวกับความเป็นอิสระดังนี้

ทฤษฎี ตัวแปรเชิงสุ่ม 2 มิติ (X,Y) จะเป็นอิสระกันก็ต่อเมื่อ

$$f(x,y) = f(x)\,f(y)$$

แต่ถ้า (X,Y) ไม่เป็นอิสระกันแล้ว เราจะเรียกว่าพึ่งพิงกันในเชิงสถิติ (Statistically Dependent)

ตัวอย่าง ถ้า X และ Y เป็นตัวแปรเชิงสุ่ม 2 มิติ ที่แทนจำนวนครั้งที่ดื่มน้ำในตอนเช้าและตอนบ่ายตามลำดับ และมีการแจกแจงร่วมดังนี้

$x \backslash y$	1	2	3	$f(x)$
1	.10	.20	.20	.5
2	.06	.12	.12	.3
3	.04	.08	.08	.2
$f(y)$	.2	.4	.4	1.0

เราจะเห็นว่าทุก ๆ (x,y) เราจะได้

$$f(x,y) = f(x)\,f(y)$$

เช่น $f(x=1, y=1) = f(x=1)\, f(y=1)$ หรือ $f(1,1) = f(1)\, f(1)$

นั่นคือ $.10 = (.5)(.2)$ เป็นต้น

ดังนั้น X และ Y จึงเป็นตัวแปรเชิงสุ่มที่เป็นอิสระกันในเชิงสถิติ

ตัวอย่าง ให้ X และ Y เป็นอายุของเครื่องมือทางจิตวิทยาที่เป็นอิเล็คโทรนิคส์ 2 ประเภท และมันมีการแจกแจงร่วมดังนี้

$$f(x,y) = e^{-(x+y)}\ ; \quad x>0,\ y>0$$

เราจะเห็นว่า $f(x) = e^{-x}$ และ $f(y) = e^{-y}$

และ $f(x/y) = e^{-(x+y)}/e^{-y} = e^{-x} = f(x)$

หรือ $f(y/x) = e^{-(x+y)}/e^{-x} = e^{-y} = f(y)$

นั่นคือ X และ Y เป็นตัวแปรเชิงสุ่มที่เป็นอิสระกันในเชิงสถิติ

3.11 ค่าคาดหวังของฟังก์ชั่นของสองตัวแปรเชิงสุ่ม

ถ้า $f(X,Y)$ เป็นฟังก์ชั่นของตัวแปรเชิงสุ่ม 2 มิติ (X,Y) เช่น $f(X,Y) = X+Y$, $f(X,Y) = XY$, หรือ $f(X,Y) = (X-Y)^2$ เป็นต้น แล้วเราก็สามารถหาค่าคาดหวังของมันได้ดังนิยาม

นิยาม ให้ (X,Y) เป็นตัวแปรเชิงสุ่ม 2 มิติ และ $f(X,Y)$ เป็นฟังก์ชั่นของ (X,Y) แล้วค่าคาดหวังของ $f(X,Y)$ กำหนดไว้ดังนี้

$$E\left[f(X,Y)\right] = \begin{cases} \sum_{x}\sum_{y}\left[f(x,y)\right] f(x,y) \\ \int_{x}\int_{y}\left[f(x,y)\right] f(x,y)\,dx\,dy \end{cases}$$

ตัวอย่าง ให้ (X,Y) แทนตัวแปรเชิงสุ่ม 2 มิติที่แทนรายได้ของสามีและภรรยาตามลำดับ โดยที่มีการแจกแจงร่วมดังนี้

$$f(x,y) = 1/15\ ; \quad \begin{matrix} 10 < x < 15 \\ 8 < y < 11 \end{matrix}$$

ถ้ากำหนดฟังก์ชั่นของ (X,Y) เป็น $f(X,Y) = X+Y$ แล้ว $f(X,Y)$ จะแทน

รายได้รวมของสามีและภรรยา ดังนั้นค่าคาดหวังจะเป็น

$$E(X+Y) = \int_{8}^{11}\int_{10}^{15} (x+y)\frac{1}{15}\,dx\,dy$$
$$= 22$$

และถ้าจะหาค่าคาดหวังของผลต่างของรายได้จะเป็น

$$E(X-Y) = \int_{8}^{11}\int_{10}^{15} (x-y)\frac{1}{15}\,dx\,dy$$
$$= 3$$

บางครั้งเราสนใจที่จะหาค่าคาดหวังของฟังก์ชั่นน่าจะเป็นเงื่อนไข ซึ่งเราก็สามารถได้จากนิยามต่อไปนี้

นิยาม ค่าคาดหวังเงื่อนไขของตัวแปรเชิงสุ่ม X เมื่อกำหนด $Y=y$ หรือ $E(X/Y=y)$ กำหนดไว้ดังนี้

$$E(X/Y=y) = \begin{cases} \sum_{x} x f(x/y) \\ \int_{x} x f(x/y)\,dx \end{cases}$$

ตัวอย่าง ให้ (X,Y) เป็นตัวแปรเชิงสุ่ม 2 มิติ ซึ่งแทนจำนวนบุตรในครอบครัวและระดับการศึกษาของหัวหน้าครอบครัว และ (X,Y) มีการแจกแจงร่วมดังนี้

$x \backslash y$	1	2	3	4	$g(x)$
1	0	.02	.05	.1	.17
2	.1	.08	.1	.1	.38
3	.3	.1	.05	0	.45
$h(y)$	.4	.2	.2	.2	1.00

ถ้าเราต้องการหาค่าคาดหวังของจำนวนบุตรในครอบครัวที่หัวหน้าครอบครัวมีระดับการศึกษาต่าง ๆ หรือ $E(X/y)$ เราจะหาได้ดังนี้

$$E(X/y) = \sum_{x} x f(x/y)$$

$$= \begin{cases} 1(0)+2(.25)+3(.75) = 2.75 & y=1 \\ 1(.1)+2(.4)+3(.5) = 2.40 & y=2 \\ 1(.25)+2(.5)+3(.25) = 2.00 & y=3 \\ 1(.5)+2(.5)+3(0) = 1.50 & y=4 \end{cases}$$

จากนิยามและตัวอย่างนี้เราจะเห็นได้ $E(X/y)$ เป็นฟังก์ชั่นของ y นั่นคือ $E(X/y) = g(y)$

และทำนองเดียวกัน $E(y/x)$ จะเป็นฟังก์ชั่นของ x ดังนั้นเราจะหาค่าคาดหวังของ $E(x/y)$ หรือ $E(y/x)$ ได้ ดังทฤษฎีต่อไปนี้

ทฤษฎี ถ้าให้ (x,y) เป็นตัวแปรเชิงสุ่ม 2 มิติ แล้ว

$$E\left[E(y/x)\right] = E(x)$$

และ

$$E\left[E(x/y)\right] = E(y)$$

จากตัวอย่างข้างบนเราหา $E\left[E(x/y)\right]$ ได้เป็น

$$\begin{aligned}
E\left[E(x/y)\right] &= E\left[f(y)\right] = \sum_{y}\left[f(y)\right]f(y) \\
&= 2.75(.4) + 2.40(.2) + 2.00(.2) \\
&\quad + 1.50(.2) \\
&= 2.28
\end{aligned}$$

แต่

$$\begin{aligned}
E(x) &= \sum_{x} x\,g(x) \\
&= 1(.17) + 2(.38) + 3(.45) \\
&= 2.28
\end{aligned}$$

นั่นคือ $E\left[E(y/x)\right] = E(x)$ ดังทฤษฎี.

3.12 มาตราวัดความสัมพันธ์ระหว่างตัวแปรเชิงสุ่ม

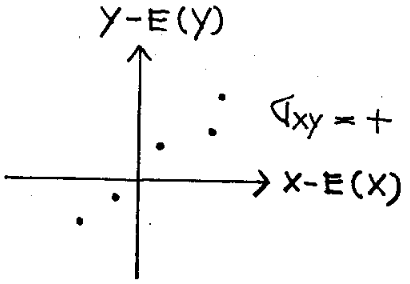
ก. **ความแปรปรวนร่วม** (Covariance) สำหรับตัวแปรเชิงสุ่มสองมิติ (x,y) นั้นเราไม่เพียงแต่ต้องการทราบว่ามันพึ่งพิงกันหรือไม่ แต่เรายังต้องการทราบว่ามันพึ่งพิงกันขนาดไหน (strength of their dependence) มาตราวัดที่ใช้ชี้ขนาดของการพึ่งพิงเชิงเส้น (linear dependence) เรียกว่า **ความแปรปรวนร่วม** ระหว่างสองตัวแปรเชิงสุ่ม

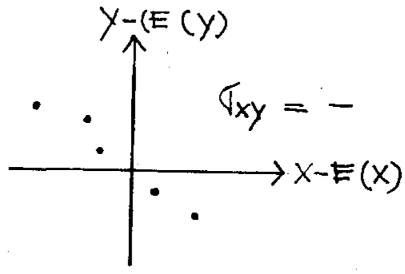
ดังนั้นจึงกล่าวได้ว่าความแปรปรวนร่วมเป็นมาตราวัดดีกรีของความสัมพันธ์เชิงเส้นระหว่างตัวแปรเชิงสุ่มทั้งสอง นั่นก็คือเป็นดัชนี (indicator) ถึงดีกรีที่สองตัวแปรเชิงสุ่มเคลื่อนไปด้วยกันในแบบเชิงเส้น

มาตราวัดนี้จะเป็นอิสระกับจุดกำเนิด เราจึงย้ายจุดกำเนิด (0,0) ไปไว้ที่ $\left[E(x), E(y)\right]$ ได้ และเราจะได้ตัวแปรใหม่เป็น $x-E(x)$ กับ $y-E(y)$ ค่าคาดหวังของผลคูณระหว่างตัวแปรใหม่นี้แหละคือความแปรปรวนร่วม ดังนิยาม

นิยาม ถ้า (x,y) เป็นตัวแปรเชิงสุ่ม 2 มิติ แล้วความแปรปรวนร่วมของ x และ y, $\mathrm{Cov}(x,y)$ กำหนดไว้ดังนี้

$$\sigma_{xy} = Cov(x,y) = E[X-E(x)][y-E(y)]$$
$$= E(XY) - E(X)E(Y)$$





ถ้า x มีแนวโน้มไปทางมาก เมื่อ y มาก และไปทางน้อยเมื่อ y น้อย แล้วจุด (x,y) จะอยู่ในที่ 1 หรือ 3 ซึ่ง σ_{xy} จะเป็นบวก (ตามรูป) แต่ถ้า x มีแนวโน้มไปทางมาก เมื่อ y น้อย และไปทางน้อยเมื่อ y มาก จุด (x,y) จะอยู่ในส่วนที่ 2 หรือ 4 แล้ว σ_{xy} จะเป็นลบ ถ้าไม่มีแนวในปีที่ x และ y จะเคลื่อนไปด้วยกันแล้ว σ_{xy} จะใกล้ศูนย์

สำหรับความแปรปรวนร่วมนี้มีค่าที่เป็นไปได้อยู่ในช่วง $(-\infty, +\infty)$ ถ้ามีค่ามาก ($+\infty$ หรือ $-\infty$) ก็แสดงว่าตัวแปรทั้งสองมีความพึ่งพิงในเชิงสถิติกันมาก และถ้าใกล้ศูนย์ก็แสดงว่ามีความพึ่งพิงเชิงสถิติกันน้อย

จากความจริงที่ว่าตัวแปรเชิงสุ่ม X และ y ที่เป็นอิสระกันนั้นเราได้

$$E(xy) = E(x)E(y)$$

นั่นก็คือ σ_{xy} จะเป็นศูนย์ด้วย ดังทฤษฎี

ทฤษฎี ให้ (X,Y) เป็นตัวแปรเชิงสุ่ม 2 มิติที่มีการแจกแจงร่วมกัน ถ้า X และ Y เป็นอิสระกันในเชิงสถิติแล้ว $Cov(X,Y) = 0$

บทกลับของทฤษฎีนี้ไม่เป็นจริงเสมอไป นั่นคือ ถ้าความแปรปรวนร่วมระหว่างสองตัวแปรเป็นศูนย์ แล้วสองตัวแปรจะเป็นอิสระกันหรือพึ่งพิงกันก็ได้

ตัวอย่าง ฟังก์ชั่นน่าจะเป็นร่วมของสองตัวแปรเชิงสุ่ม (X,Y) กำหนดไว้ดังนี้

$x \backslash y$	1	2	3	$g(x)$
1	0	1/3	0	1/3
2	1/3	0	1/3	2/3
$h(y)$	1/3	1/3	1/3	1.00

$$E(X) = 1(1/3) + 2(2/3) = 5/3$$
$$E(Y) = 1(1/3) + 2(1/3) + 3(1/3) = 2$$
$$E(XY) = 1(1)(0) + 1(2)(1/3) + 1(3)(0) + 2(1)(1/3) + 2(2)(0) + 2(3)(1/3) = 10/3$$

ดังนั้น $\quad Cov(X,Y) = E(XY) - E(X)E(Y) = 10/3 - (5/3)(2) = 0$

อย่างไรก็ตามถึงแม้ว่า $Cov(X,Y) = 0$ ตัวแปรเชิงสุ่ม X และ Y ก็ยังพึ่งพิงกัน เพราะ $f(x,y) \neq f(x)f(y)$

กฎเกี่ยวกับความแปรปรวนร่วม (Some Properties of Covariance)

ให้ X และ Y เป็นตัวแปรเชิงสุ่มที่มีการแจกแจงร่วมกัน และ a, b เป็นค่าคงที่ แล้วเราจะได้กฎที่น่าสนใจต่อไปนี้

(1) $Cov(X+a, Y+b) = Cov(X,Y)$

(2) $Cov(aX, bY) = ab\, Cov(X,Y)$

(3) $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab\, Cov(X,Y)$

$\quad = a^2V(X) + b^2V(Y)$ ถ้า X และ Y เป็นอิสระกัน

ความแปรปรวนร่วมนี้จะมีหน่วยเป็นผลคูณของหน่วยทั้งสองของสองตัวแปร บางครั้งหน่วยก็แตกต่างกันมาก เช่น X มีหน่วยเป็นฟุต Y มีหน่วยเป็นบาท จึงทำให้ยุ่งยากในการแปลความหมาย และบางทีเราต้องการทราบว่าตัวแปรเชิงสุ่ม 2 คู่ เช่น (X_1, Y_1) และ (X_2, Y_2) นั้นคู่ไหนความสัมพันธ์เชิงเส้นกันมากกว่า เราจะประสบความยุ่งยากอีกถ้าขนาดของค่าของตัวแปรทั้งสองคู่แตกต่างกันมาก เพื่อขจัดปัญหาเหล่านี้ นักสถิติจึงได้แปลงตัวแปรทั้งสองให้เป็นหน่วยมาตรฐาน (Standard Unit) โดยเอาส่วนเบี่ยงเบนมาตรฐานไปหารส่วนเบี่ยงเบนของตัวแปรนั้นเอง แล้วเราจะได้มาตราวัดความสัมพันธ์เชิงเส้นใหม่ ซึ่งให้ชื่อว่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient, f) ซึ่งกำหนดไว้ดังนี้

นิยาม ถ้า (X,Y) เป็นตัวแปรเชิงสุ่ม 2 มิติ ที่มีการแจกแจงร่วมกัน แล้วสัมประสิทธิ์

สหสัมพันธ์ระหว่างตัวแปรทั้งสองจะกำหนดไว้ว่า

$$\rho = E\left(\frac{X-E(X)}{\sigma_x}\right)\left(\frac{Y-E(Y)}{\sigma_y}\right)$$

$$= \frac{Cov(X,y)}{\sigma_x \sigma_y}$$

สัมประสิทธิสหสัมพันธ์นี้จะมีค่าอยู่ระหว่าง -1 กับ 1 นั่นคือ

$$-1 < \rho < +1$$

ถ้าตัวแปรทั้งสองมีความผันแปรร่วมอย่างสมบูรณ์ (Perfect Covariability) แล้ว ρ = -1 หรือ +1 แล้วแต่ว่าจะผันแปรไปคนละทาง หรือทางเดียวกัน และถ้าดีกรีของความผันแปรน้อยกว่าก็จะให้สัมประสิทธิ์อยู่ระหว่าง -1 กับ +1 หรือ

สำหรับตัวแปรเชิงสุ่ม X และ Y ที่เป็นอิสระกันนั้น $Cov(X,Y)=0$ และจะให้ $\rho = 0$ ด้วย

เรามีทฤษฎีเกี่ยวกับสัมประสิทธิสหสัมพันธ์ที่น่าสนใจอีกดังนี้

ทฤษฎี ถ้า X และ Y แต่ก็เป็นตัวแปรเชิงสุ่มที่มีความสัมพันธ์กันดังนี้

$$Y = a+bX$$

ในเมื่อ a, b เป็นค่าคงที่

แล้ว $\rho^2 = 1, \rho = +1$ ถ้า $b > 0$, และ $\rho = -1$ ถ้า $b < 0$

ทฤษฎี ถ้า ρ_{xy} เป็นสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเชิงสุ่ม X และ Y และถ้า U และ V เป็นฟังก์ชั่นเชิงเส้นของ X และ Y ดังนี้

$$U = a+bx, \quad V = c+dy;$$

$$\rho_{UV} = \frac{bd}{|bd|}\rho_{xy}$$

และถ้า b และ d ต่างก็มากกว่าศูนย์ หรือต่างก็น้อยกว่าศูนย์ แล้ว $\rho_{UV} = \rho_{xy}$

ตัวอย่าง ให้ X และ Y เป็นตัวแปรเชิงสุ่มที่แสดงถึงการเปลี่ยนแปลงอารมณ์ของสามีและภรรยาตามลำดับ และทั้งสองมีการแจกแจงร่วมดังนี้

$x \backslash y$	-2	2	$f(x)$
-2	.2	.05	.25
2	.15	.6	.75
$f(y)$	.35	.65	1.00

เราจะเห็นว่าตัวแปรเชิงสุ่มทั้งสองนี้พึ่งพิงกันในเชิงสถิติ เพราะว่าทุก ๆ (x,y) เราจะได้ว่า $f(x,y) \neq f(x)\,f(y)$

ดีกรีของความพึ่งพิงเชิงเส้นจะหาได้ดังนี้

$$
\begin{aligned}
E(X) &= -2(.25)+2(.75) = 1.00 \\
&= -2(.35)+2(.65) = 0.60 \\
&= -2(-2)(-2)+(-2)(2)(.05) \\
&\quad +2(-2)(.15)+2(2)(.6) \\
&= 2.4
\end{aligned}
$$

ดังนั้น $Cov(x,y) = 2.4-1.00\,(.60)$

$$= 1.8$$

$$V(x) = (-2-1)^2\,(.25)+(2-1)^2\,(.75) = 3.00$$

$$V(y) = (-2-.6)^2\,(.35)+(2-.6)^2\,(.65) = 3.64$$

เพราะฉะนั้น $\rho = \dfrac{2.4}{\sqrt{3(3.64)}} \cong .77$

3.13 การแจกแจงน่าจะเป็นแบบมาตรฐานบางชนิด (Some Special Probability Distributions)

ในการทดลองเชิงสุ่มใด ๆ จะให้การแจกแจงของตัวแปรเชิงสุ่มออกมา บางครั้งการแจกแจงจะมีรูปฟอร์มเป็นแบบทั่ว ๆ ไปซึ่งเราเรียกว่าการแจกแจงน่าจะเป็นแบบมาตรฐานซึ่งมีอยู่หลายชนิด ที่น่าสนใจก็มีดังต่อไปนี้

ก. กรณีตัวแปรเชิงสุ่มแบบไม่ต่อเนื่อง

ก.1 การแจกแจงยูนิฟอร์ม (Uniform distributions) การทดลองเชิงสุ่มที่มีผลทดลองที่เป็นไปได้ n วิธี โดยที่แต่ละวิธีไม่เกิดร่วมกัน แต่มีโอกาสที่จะเกิดขึ้นเท่า ๆ กัน ตัวแปรเชิงสุ่ม X ที่กำหนดจากกลุ่มผลทดลองของการทดลองแบบนี้ จะมีค่าต่าง ๆ กัน n ค่า คือ $x_1, x_2, \ldots, x_n$ และจะมีการแจกแจงยูนิฟอร์มดังนี้

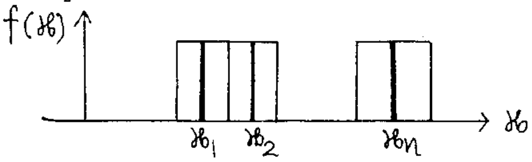
$$f(x) = 1/n \;;\; x = x_1, x_2, \ldots, x_n$$

ค่าคาดหวังและความแปรปรวนของการแจกแจงนี้จะเป็น

$$E(X) = \sum_{i=1}^{n} x_i / n$$

$$V(X) = \sum_i [x_i - E(x)]^2 / n$$

กราฟของการแจกแจงยูนิฟอร์มจะมีลักษณะดังนี้



เนื่องจากกราฟของการแจกแจงนี้มีลักษณะเป็นสี่เหลี่ยมผืนผ้า เราจึงมักเรียกชื่ออีกอย่างว่า การแจกแจงสี่เหลี่ยม (Rectangular distribution)

การทดลองเชิงสุ่มที่ให้การแจกแจงยูนิฟอร์มนี้มีมากมาย เช่นสังเกตหน้าของลูกเต๋าจากการโยน 1 ครั้ง เป็นต้น

ตัวอย่าง วงล้ออันหนึ่งแบ่งเป็นสิบส่วนเท่า ๆ กัน และมีหมายเลข 0 ถึง 9 กำกับอยู่ ถ้าหมุนวงล้อนี้ และให้ **X** เป็นหมายเลขที่วงล้อหยุดตรงลูกศรแล้ว X จะมีการแจกแจงยูนิฟอร์มดังนี้

$$f(x) = 1/10 \ ; \ x = 0,1,2,\ldots\ldots,9 \ .$$

ก.2 การแจกแจงทวินาม (Binomial distribution) ถ้าการทดลองเชิงสุ่มอย่างหนึ่งมีคุณสมบัติดังนี้

การทดลองจะกระทำซ้ำ ๆ กัน (Repeated trials) เป็นจำนวนจำกัดครั้งคือ n โดยกระทำภายใต้สภาวะการณ์เดียวกัน

ผลทดลองที่เกิดขึ้นจากการทดลองแต่ละครั้งจะเป็นประเภทหนึ่งใน 2 ประเภท คือ "ประสบความสำเร็จ" หรือ "ล้มเหลว"

ความน่าจะเป็นที่จะประสบความสำเร็จในการทดลองหนึ่ง จะคงที่ตลอดทุก ๆ ครั้งของการทดลอง และจะให้เป็น p

การทดลองแต่ละครั้งจะเป็นอิสระกัน

ถ้าให้ X เป็นตัวแปรเชิงสุ่มที่แทนจำนวนครั้งที่ประสบความสำเร็จในการทดลอง n ครั้งนี้ แล้วเราจะเรียก X ว่าตัวแปรเชิงสุ่มทวินาม (Binomial Random Variable) ซึ่งมีการแจกแจงทวินามดังนี้

$$f(x) = \binom{n}{x} p^{x} q^{n-x} \ ; \ \begin{array}{l} x = 0,1,2,\ldots,n \\ q = 1-p \end{array}$$

ค่าคาดหวังและความแปรปรวนของตัวแปรเชิงสุ่มทวินามเป็นดังนี้

$$E(X) = np$$

$$V(X) = npq$$

ตัวอย่าง ในการสอบไล่คราวหนึ่ง นายรัฐวีร์ทำข้อทดสอบไปได้ 95 ข้อ เผอิญหมดเวลาก่อนเขาจึงเดาอีก 5 ข้อที่เหลือ ข้อทดสอบนี้เป็นปรนัยที่มีข้อเลือก 5 ข้อ

ถ้าให้ X เป็นจำนวนข้อที่เดาถูก แล้ว X จะมีการแจกแจงทวินาที่มี $n=5, p=1/5=.20$ ดังนี้

$$f(x) = \binom{5}{x}(.20)^{x}(.80)^{5-x}; \quad x = 0,1,2,\ldots,5$$

โอกาสที่เดาไม่ถูกเลยจะเป็น $f(0) = \binom{5}{0}(.20)^0(.80)^{5-0} = (.80)^5$

โอกาสที่เดาถูกอย่างน้อย 4 ข้อ จะเป็น $f(4) + f(5)$

$$= \binom{5}{4}(.20)^4(.80)^{5-4} + \binom{5}{5}(.20)^5(.80)^{5-5}$$

$$= 5(.20)^4(.80) + (.20)^5 = (.20)^4(4+.20)$$

$$= (4.20)(.20)^4$$

โดยเฉลี่ยแล้วเขาจะเดาถูกเท่ากับ $np = 5(.20) = 1$

และมีความแปรปรวนของจำนวนข้อที่เดาถูกเท่ากับ $npq = 5(.20)(.80)$

$= 0.80$

(ก) **ฟังก์ชั่นแจกแจงสะสมแบบทวินาม** (Binomial distribution) ฟังก์ชั่นแจกแจงทวินามจะให้ความน่าจะเป็นที่จะประสบความสำเร็จ x ครั้ง หรือน้อยกว่าจากการทดลอง n ครั้ง นั่นคือ

$$F(x) = P(X \le x)$$
$$= \sum_{s=0}^{x} \binom{n}{s} p^{s n-s}$$

สำหรับความน่าจะเป็นที่จะประสบความสำเร็จอย่างน้อย x ครั้ง หรือ $P(X \ge x)$ กำหนดไว้ดังนี้

$$P(X \ge x) = \sum_{s=x}^{n} \binom{n}{s} p^s q^{n-s}$$
$$= 1 - P(X \le x-1) = 1 - F(x-1)$$

ดังนั้น $P(X \le x)$ สามารถเขียนใหม่ได้เป็น

$$P(X \le x) = 1 - P(X \ge x+1)$$

จากตัวอย่างของการเดาข้อสอบนั้น ความน่าจะเป็นที่จะเดาถูกไม่เกิน 4 ข้อ จะหาได้ดังนี้

$$P(X \le 4) = \sum_{s=0}^{4} \binom{5}{s}(.20)^s(.80)^{5-s}$$

$$= f(0) + f(1) + f(2) + f(3) + f(4)$$

หรือ $$P(X \le 4) = 1 - P(X \ge 5) = 1 - f(5)$$
$$= 1 - (.20)^5$$

ถ้าจะหาโอกาสที่จะเดาถูกอย่างน้อย 1 ข้อ ก็หาได้ดังนี้

$$P(X \geq 1) = \sum_{x=1}^{5} \binom{5}{x} p (.20)^{x} (.80)^{5-x}$$
$$= f(1) + f(2) + f(3) + f(4) + f(5)$$

หรือ $$P(X \geq 1) = 1 - P(X \leq 0) = 1 - f(0)$$
$$= 1 - (.80)^5$$

ในการหาความน่าจะเป็นในแต่ละค่าของตัวแปรเชิงสุ่มทวินามหรือหาความน่าจะเป็นสะสม จะหาได้จากตารางแจกแจงทวินามในหนังสือสถิติ หรือทฤษฎีความน่าจะเป็นทั่ว ๆ ไป ได้ เช่นถ้าทราบค่า n และ p ว่าเป็น 5 และ .20 ตามลำดับ แล้วเราจะหาค่าต่าง ๆ จากตารางทวินามได้ดังนี้

$$f(0) = .3277 \quad f(1) = .4096 \quad f(2) = .2048$$
$$f(3) = .0512 \quad f(4) = .0064 \quad f(5) = .0003$$

$$P(X \geq 1) = .6723 \quad P(X \geq 2) = .2627 \quad P(X \leq 3) = .9933$$
$$P(X \leq 4) = .9997$$

(ข) การแจกแจงทวินามมีความสัมพันธ์กับการแจกแจงเอฟ (F) ซึ่งจะได้กล่าวต่อไป

ก.3 การแจกแจงมหุนาม (Multinomial distribution) การแจกแนงนี้เป็นแบบทั่วไปของการแจกแจงทวินาม นั่นคือเกี่ยวข้องกับการทดลองที่ให้ผลทดลองแบ่งเป็นประเภทมากกว่า 2 ประเภท และการทดลองนั้นมีคุณสมบัติดังนี้

(1) การทดลองแต่ละครั้งจะให้ผลทดลองเป็นประเภทหนึ่งใน k ประเภท$(C_1, C_2, \ldots, C_k)$

(2) ความน่าจะเป็นที่ประเภทต่าง ๆ เหล่านี้จะเกิดขึ้นให้เท่ากับ $p_1, p_2, \ldots$ ตามลำดับ และความน่าจะเป็นนี้จะคงที่เสมอสำหรับการทดลองครั้งใด ๆ

(3) การทดลองครั้งหนึ่ง ๆ จะเป็นอิสระกับการทดลองครั้งอื่น ๆ ทั้งหมด

(4) การทดลองต้องกระทำซ้ำ ๆ กันจำนวนจำกัดคือ n และต้องทำในสภาวะการณ์เดียวกัน

ถ้าให้ $X_1, X_2, \ldots, X_k$ เป็นจำนวนครั้งที่ $C_1, C_2, \ldots, C_k$ เกิดขึ้นในการทดลอง n ครั้ง แล้ว $X_1, X_2, \ldots, X_k$ จะเป็นตัวแปรเชิงสุ่มแบบทวินามที่มีการแจกแจงดังนี้

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{(x_1!)(x_2!)\ldots(x_k!)} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k};$$
$$x_i = 0, 1, 2, \ldots, n \quad (i = 1, 2, \ldots, k)$$
$$\sum x_i = n, \quad \sum p_i = 1$$

ในการแจกแจงพหุนามนั้นจำนวนครั้งที่เกิดขึ้นของ E_i $(i=1,2,\ldots,k)$ ใด ๆ ในการทดลองซ้ำ ๆ กัน n ครั้ง ที่เป็นอิสระกันนั้นอาจจะถือว่าเป็นตัวแปรเชิงสุ่มทวินามตัวหนึ่งได้ โดยมีค่าคาดหวังและความแปรปรวนดังนี้

$$E(X_i) = np_i \qquad i = 1,2,\ldots,k$$
$$V(X_i) = np_iq_i$$

ตัวอย่าง นักจิตวิทยาได้ทำการทดลองกับหนู 5 ตัว โดยให้หนูแต่ละตัวเดินทางเข้าไปในกรงที่มีอาหาร ทางที่จะเข้าไปมีอยู่ 5 ทาง ซึ่งเหมือนกันแต่ทาสีต่างกัน คือ แดง, เขียว, น้ำเงิน, ดำ และขาว และหนูแต่ละตัวจะมีโอกาสที่จะเลือกทางใดทางหนึ่งเท่า ๆ กัน

ถ้าให้ $x_1, x_2, \ldots, x_k$ เป็นจำนวนหนูที่อยู่ในทางสีแดง, เขียว, น้ำเงิน ดำ และขาว ตามลำดับ แล้ว $x_1, x_2, \ldots, x_k$ เป็นตัวแปรเชิงสุ่มพหุนามที่มีการแจกแจงดังนี้

$$f(x_1, x_2, \ldots, x_5) = \frac{5!}{(x_1!)(x_2!)\cdots(x_5!)}(1/5)^{x_1}(1/5)^{x_2}\cdots(1/5)^{x_5};$$
$$x_i = 0,1,2,\ldots,5 \quad (i = 1,2,\ldots,5)$$
$$\sum x_i = 5, \quad \sum p_i = 1$$

โอกาสที่ทางทั้ง 5 ทางนั้นจะมีหนูอยู่ทางละตัวจะเป็น $P(X_1 = 1, X_2 = 1, \ldots, X_5 = 1)$

$$= f(1,1,1,1,1) = \frac{5!}{1!\,1!\cdots 1!}(1/5)^1(1/5)^1\cdots(1/5)^1$$
$$= 120(1/5)^5$$

โอกาสที่ทางสีแดงจะมีหนู 3 ตัว ทางสีเขียวและน้ำเงินมีหนูอย่างละตัวจะเป็น $f(3, 1, 1, 0, 0) =$

$$\frac{5!}{(3!)(1!)\ldots(0!)} \quad 1/5)^3(1/5)^1\ldots(1/5)^0$$
$$= 20(1/5)^5$$

โอกาสที่ทางสีดำจะมีหนู 4 ตัว เท่ากับ $P(X_4 = 4)$

$$= \binom{5}{4}(1/5)^4(4/5)^1 = 4(1/5)^4$$

ก.4 การแจกแจงเรขาคณิต (Geometric distribution) ในการทดลองเชิงสุ่มแบบทวินามนั้น ถ้าเราสนใจ "จำนวนครั้งที่ทำการทดลองจนกว่าจะประสบความสำเร็จ" แล้วตัวแปรเชิงสุ่ม X ซึ่งแทนจำนวนครั้งที่ทำการทดลองนั้นจะเป็นตัวแปรเชิงสุ่มแบบเรขาคณิตที่มีการแจกแจงดังนี้

$$f(x) = pq^{x-1} \quad ; \quad x = 1,2,3,\ldots$$

และมีค่าคาดหวังและความแปรปรวนดังนี้

$$E(x) = 1/p$$
$$V(x) = q/p^2$$

ตัวอย่าง นักจิตวิทยาได้ทำการทดลองอย่างหนึ่งในห้องทดลอง ค่าใช้จ่ายในการทดลองครั้งหนึ่งเป็นเงิน 20000 บาท แต่ถ้าผลการทดลองล้มเหลวต้องเพิ่มค่าใช้จ่ายอีก 5000 บาท เพราะต้องเปลี่ยนแปลงบางส่วนก่อนที่จะทำการทดลองครั้งต่อไป

สำหรับความน่าจะเป็นที่จะประสบความสำเร็จในการทดลองครั้งหนึ่งครั้งใดเป็น 0.8 และการทดลองแต่ละครั้งเป็นอิสระกัน

จงหาการแจงน่าจะเป็นของจำนวนครั้งที่ทำการทดลอง

ถ้าทำการทดลองจนกว่าจะประสบความสำเร็จจะต้องเสียค่าใช้จ่ายเฉลี่ยในการทดลองทั้งหมดเท่าใด?

ถ้าให้ X เป็นจำนวนครั้งที่จะต้องทำจนกว่าจะประสบความสำเร็จ แล้ว X จะเป็นตัวแปรเชิงสุ่มแบบเรขาคณิตที่มีการแจกแจงดังนี้

$$f(x) = (.80)\,(.20)^{x-1} \;;\; x = 1, 2, 3, \ldots$$

โอกาสที่จะทำการทดลอง 3 ครั้งแล้วประสบความสำเร็จจะเป็น $P(X = 3) = f(2) = (.80)(.20)^{3-1} = .032$

ถ้าให้ C เป็นค่าใช้จ่ายทั้งหมดของการทดลองจนกว่าประสบความสำเร็จแล้วเราจะได้ว่า

$$\begin{aligned} C &= 20000(X) + 5000\,(X-1) \\ &= 25000(X) - 5000 \end{aligned}$$

ดังนั้น

$$\begin{aligned} E(C) &= 25000\,E(X) - 5000 \\ &= 25000\,(1/.8) - 5000 \\ &= 31250 - 5000 = 26520 \text{ บาท} \end{aligned}$$

ก.5 การแจกแจงทวินามนิเสธ (Negative Binomial Distribution) ในการทดลองเชิงสุ่มเช่นเดียวกับการทดลองแบบเรขาคณิต ถ้าต้องการประสบความสำเร็จ k^1 ครั้ง จึงจะเลิกทำการทดลอง แล้วจำนวนครั้งที่ต้องทำการทดลองจะเป็นตัวแปรเชิงสุ่มแบบทวินามนิเสธ ซึ่งมีการแจกแจงดังนี้

$$f(x) = \binom{x-1}{k-1} p^k q^{x-k} \quad ; \quad x = k, k+1, \ldots$$

และมีค่าคาดหวังและความแปรปรวนดังนี้

$E(X) = k/p$

$V(X) = kq/p^2$

การแจกแจงแบบนี้มีชื่อเรียกอีกว่า การแจกแจงปาสคาล (Pascal distribution)

การแจกแจงทวินามนิเสธนี้มีความสัมพันธ์กับการแจกแจงทวินาม ดังนี้

$$(1)\ P(X \leq r) = P(y \geq k) \Rightarrow \sum_{x=k}^{r} \binom{x-1}{k-1} p^k q^{x-k} = \sum_{y=k}^{r} \binom{r}{y} p^y q^r$$

$$(2)\ P(X > r) = P(y < k)$$

ในเมื่อ X มีการแจกแจงทวินามนิเสธที่มีพารามิเตอร์ k และ p

y มีการแจกแจงทวินามที่มีพารามิเตอร์ r, p

ดังนั้นในการคำนวนความน่าจะเป็นของการแจกแจงทวินามนิเสธ เราจึงอาศัยการแจกแจงทวินามได้อีกด้วย

ตัวอย่าง พนักงานขายที่ได้รับการอบรมอย่างดีจากนักจิตวิทยาได้เสนอขายของให้แก่ลูกค้าจำนวนหนึ่ง เป็นที่ทราบกันว่าลูกค้าจะสั่งของจากพนักงานขายประเภทนี้ถึง 40% ถ้าพนักงานขายต้องการให้ลูกค้า 3 รายสั่งของ โอกาสที่เขาจะต้องติดต่อลูกค้าเพียง 5 ราย เท่านั้น จะเป็นเท่าใด?

ให้ X เป็นจำนวนลูกค้าที่พนักงานต้องเสนอขาย แล้ว X จะมีการแจกแจงดังนี้

$$f(x) = \binom{x-1}{3-1} (.4)^3 (.6)^{x-3} \quad ; \quad x = 3, 4, \ldots$$

ดังนั้น

$$P(X=5) = \binom{5-1}{3-1} (.4)^3 (.6)^{5-3} = 13824$$

สำหรับ E(X) และ V(X) หาได้ดังนี้

$E(X) = 3/.4 = 7.5$

$V(X) = 3(.6)/(.4)^2 = 10.25$

ถ้าต้องการหาโอกาสที่จะต้องติดต่อลูกค้าไม่เกิน 10 ราย นั้นจะทำได้ดังนี้

$$\begin{aligned} P(X \leq 10) &= P(y \geq 3) = \sum_{y=5}^{10} \binom{10}{y} (.4)^y (.6)^{10-y} \\ &= 1 - \sum_{y=0}^{4} \binom{10}{y} (.4)^y (.6)^{10-y} \\ &= 1 - .6320 = .3680 \end{aligned}$$

ก.6 **การแจกแจงไฮเพอร์จีโอเมทริค (Hypergeometric distribution)** ในการแจกแจงทวินามนั้นเกี่ยวข้องการทดลองเชิงสุ่มซ้ำ ๆ ที่เป็นอิสระกัน แต่ถ้าไม่เป็นอิสระกัน และความน่าจะเป็นที่จะประสบความสำเร็จครั้งใด ๆ ไม่คงที่ แล้วการทดลองแบบนี้จะเกี่ยวข้องการการแจกแจงไฮเปอร์จีโอเมทริค

โดยทั่วไปการแจกแจงนี้เกี่ยวข้องกับตัวอย่างสุ่มขนาด n หน่วย ที่เลือกสุ่มจากประชากรจำกัดที่มีขนาด N และจำนวนหน่วยทั้งหมดในประชากรจะแยกออกได้ 2 พวก คือ พวกแรกเรียกว่า "ประสบความสำเร็จ" ซึ่งมี m หน่วย พวกสองเรียกว่า "ล้มเหลว" จะมี (N-m) หน่วย ถ้าให้ X เป็นจำนวนหน่วยของพวกแรกในตัวอย่างสุ่มขนาด n หน่วย แล้ว X จะมีการแจกแจงดังนี้

$$f(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \quad ; \max(0, n-(N-m)) \leq x \leq \min(n, m)$$

ในเมื่อ max(0, n− (N−m)) เป็นจำนวนเลขที่เลือกเอาค่ามากระหว่าง 0 กับ n−(N−m) เช่น man(0, 5) = 5 และ man(0, − 3) = 0 เป็นต้น

min (n, m) เป็นจำนวนเลขที่เลือกเอาค่าน้อยระหว่าง n กับ m เช่น min(5, 10) = 5 และ min (7, 3) = 3 เป็นต้น

ค่าคาดหวังและความแปรปรวนของตัวแปรเชิงสุ่มนี้จะเป็น

$E(X) = n(m/N)$

$V(X) = n\,(m/N)\,(1-m/N)\,\frac{N-n}{N-1}$

จะเห็นได้ว่าค่าคาดหวังของการแจกแจงไฮเปอร์จีโอเมทริค จะเท่ากับของแบบทวินาม (เพราะ $m/N = p$ นั่นเอง) แต่ความแปรปรวนต่างกันด้วยตัวคูณ $\frac{N-n}{N-1}$ ถ้า N มากพอ ($N \rightarrow \infty$) แล้วตัวคูณ $\frac{N-n}{N-1}$ จะเข้าใกล้ 1 มาก ซึ่งจะได้ความแปรปรวนเท่ากัน

ในกรณีที่ N มากพอการแจกแจงไฮเปอร์จีโอเมทริค จะมีความโน้มเอียงที่จะเท่ากับการแจกแจงทวินาม นั่นคือ

$$\lim_{N\to\infty} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \longrightarrow \binom{n}{x} p^{x} q^{n-x} \quad ; \quad p = m/N$$

การประยุกต์ การแจกแจงนี้ได้นำไปใช้เป็นประโยชน์ในการประมาณจำนวนหน่วยในประชากรที่เราไม่สามารถแจงนับทั้งหมดได้ เช่น สัตว์ในป่า ปลาในน้ำ และยังใช้ประมาณ

จำนวนหน่วยที่เป็นลักษระหนึ่งของประชากรได้อีก เช่น ประมาณจำนวนเครื่องมือทางจิตวิทยาที่ใช้การไม่ได้จากโรงงานที่ส่งมาให้

(1) การประมาณจำนวนหน่วยในประชากร ลองพิจารณาตัวแบบการสุ่มตัวอย่างทางตรง (Direct Sampling Model) จากปัญหาในการประมาณจำนวนปลาในทะเลสาป ซึ่งมีขบวนการดังนี้

– จับปลามาจำนวนหนึ่งจากทะเลสาป แล้วทำเครื่องหมายไว้ สมมติว่าจับมาได้ m ตัว และให้ปล่อยปลาที่ทำเครื่องหมายลงไปในทะเลสาปตามเดิม

– ต่อมาจับปลามาอีกจำนวนหนึ่ง สมมติว่า n ตัว แล้วนับจำนวนปลาที่มีเครื่องหมายไว้

ถ้าให้ X เป็นจำนวนปลาที่มีเครื่องหมายจากการจับปลาครั้งที่สอง และ N เป็นจำนวนปลาทั้งหมดที่ยังไม่ทราบค่า แล้วการแจกแจงของ X จะเป็น

$$f(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \quad ; \; x = 0, 1, 2, \ldots, min(n, m)$$

เมื่อให้ $\hat{N}$ เป็นค่าประมาณแบบน่าจะเป็นมากสุด (Maximum Likelihood Estmate) ของ N แล้วเราจะได้ว่า

$\hat{N} = [mn/x]$

นั่นคือ $\hat{N}$ เป็นจำนวนเลขเต็มบวกที่มากสุดจากความสัมพันธ์ $\hat{N} \leqslant mn/x$

ตัวอย่าง จับปลาในบ่อมาครั้งแรก 100 ตัว ทำเครื่องหมายแล้วปล่อยไป ตอนหลังจับปลามาอีก 80 ตัว ปรากฏว่ามีปลาที่มีเครื่องหมาย 2 ตัว แล้วปลาในบ่อจะมีประมาณ $\hat{N}$

$\hat{N} = [mn/x] = [100 (80)/2]$

$= 4000$ตัว

(2) การประมาณจำนวนหน่วยที่มีลักษณะหนึ่งของประชากร ลองพิจารณาปัญหาในการประมาณจำนวนนักศึกษาจิตวิทยาที่บิดาประกอบอาชีพเกษตรกรรม มีวิธีการประมาณดังนี้

ให้ N เป็นจำนวนนักศึกษาจิตวิทยาทั้งหมด n เป็นขนาดตัวอย่างของนักศึกษา xเป็นจำนวนนักศึกษาที่บิดาทำเกษตรกรรมในตัวอย่างนั้น และ m เป็นจำนวนนักศึกษาทั้งหมดที่บิดาทำเกษตรกรรม ซึ่งเราต้องการประมาณค่า ดังนั้นโอกาสที่นักศึกษาที่บิดาทำเกษตรกรรม xคนในตัวอย่างขนาด n จะเท่ากับ

$$\binom{m}{x}\binom{N-m}{n-x} \Big/ \binom{N}{n}$$

ค่าประมาณของ m คือ m ก็จะหาได้เช่นเดียวกับ $\hat{N}$ ซึ่งจะได้เป็น

$$\hat{m} = [(N + 1)x/n]$$

หรือ $\hat{m}$ เป็นจำนวนเลขเต็มบวกที่ได้จากความสัมพันธ์ $\hat{m} \leq (N + 1)x/n$

ตัวอย่าง นักจิตวิทยาได้ส่งเครื่องมือจิตวิทยามา 1000 เครื่อง ผู้ตรวจสอบได้สุ่มเครื่องมือมา 25 เครื่อง พบว่ามี 1 เครื่องที่ใช้การไม่ได้ ดังนั้นจำนวนเครื่องมือที่ใช้การไม่ได้จะมีประมาณ $\hat{m}$

$$\hat{m} = [(N + 1)x/n] = [(1000 + 1)(1)/25]$$
$$= 40 \text{ เครื่อง}$$

ในการทดลองแบบไฮเปอร์จีโอเมทริคนั้น ถ้าจำนวนหน่วยในประชากรแบ่งออกได้เป็น k พวก ($C_1, C_2, \ldots, C$) โดยแต่ละพวกมีจำนวน $m_1, m_2, \ldots, m_k$ ตามลำดับ เมื่อให้ $X_1, X_2, \ldots, X_h$ เป็นจำนวนครั้งที่เหตุการณ์ $C_1, C_2, \ldots, C_h$ เกิดขึ้น แล้ว $X_1, X_2, \ldots X_h$ จะมีการแจกแจงไฮเปอร์จีโอเมทริคทั่วไป นั่นคือ

$$f(x_1, x_2, \ldots, x_k) = \binom{m_1}{x_1}\binom{m_2}{x_2}\cdots\binom{m_k}{x_k} \Big/ \binom{N}{n}$$

$$x_i = 0, 1, 2, \ldots, n \quad (i = 1, 2, \ldots, k)$$
$$\Sigma x_i = n, \quad \Sigma m_i = N$$

ตัวอย่าง โรงงานแห่งหนึ่งมีคนงานฝ่ายผลิตอยู่ 40 คน คุมเครื่องจักร 2 คน ฝ่ายบริหาร 3 คน ฝ่ายการตลาด 5 คน พอสิ้นปีเจ้าของโรงงานให้รางวัลคนงาน 5 รางวัลโดยการจับฉลากที่มีฉลากเท่ากับคนงานทั้งหมด คือ 50 ฉลาก

ถ้าให้ X_1, X_2, X_3, X_4 เป็นจำนวนรางวัลที่ฝ่ายผลิต, เครื่องจักร, บริหาร, และ ฝ่ายตลาด ได้รับตามลำดับ แล้ว X_1, X_2, X_3, X_4 จะมีการแจกแจงดังนี้

$$f(x_1, x_2, x_3, x_4) = \binom{40}{x_1}\binom{2}{x_2}\binom{3}{x_3}\binom{5}{x_4} \Big/ \binom{50}{5}; \quad \begin{array}{l} x = 0,1,2,3,4,5 \\ (i = 1,2,3,4) \\ \Sigma x_i = 5 \end{array}$$

ดังนั้นโอกาสที่รางวัลทั้ง 5 นั้น จะเป็นของฝ่ายผลิต, เครื่องจักร, บริหาร, และตลาด เป็นจำนวน 2, 1, 1 , 1 รางวัล ตามลำดับ จะเป็นดังนี้

$$f(2,1,1,1) = \binom{40}{2}\binom{2}{1}\binom{3}{1}\binom{5}{1} \Big/ \binom{50}{5} \approx 0$$

ก.7 การแจกแจงไฮเปอร์จีโอเมทริคนิเสธ (Negative Hypergeometric distribution) การแจกแจงแบบนี้เกี่ยวข้องกับตัวอย่างสุ่มที่ดึงมาจากประชากรที่มีขนาด N หน่วย และประชากรนี้แบ่งออกได้เป็น 2 พวก คือพวกหนึ่งเรียกว่า "ประสบความสำเร็จ" จะมี m หน่วย อีกพวก

เรียกว่า "ล้มเหลว" จะมี N−m หน่วย ถ้าต้องการจะสุ่มตัวอย่างจนประสบความสำเร็จถึง k ครั้ง แล้วจำนวนครั้งหรือขนาดตัวอย่างจะเป็นตัวแปรเชิงสุ่ม จะให้เป็น X และจะมีการแจกแจงแบบไฮเปอร์จีโอเมทริคนิเสธ ดังนี้

$$f(X) = \frac{\binom{m}{k-1}\binom{N-m}{X-k}}{\binom{N}{X-1}} \cdot \frac{m-k+1}{N-X+1} \qquad X = k, k+1, \ldots$$

$$= \frac{\binom{m}{k-1}\binom{N-m}{X-k}(X-1)!}{(N)_X} \cdot (m-k+1)$$

ในเมื่อ

$$(N)_X = \frac{N!}{(N-X)!} = {}^{N}P_X$$

การแจกแจงแบบนี้นำไปใช้ในการประมาณจำนวนสัตว์ (Zoological Sqmple Census) ได้เช่นเดียวกับการแจกแจงไฮเปอร์จีโอเมทริคดังตัวแบบการสุ่มตัวอย่างชนิดผกผัน (Inverse Sampling Model) ต่อไปนี้

ปัญหาที่พิจารณาก็เช่นเดียวกับตัวแบบการสุ่มตัวอย่างทางตรง โดยมีวิธีการตามเป็นขั้น ๆ ดังนี้

– จับปลามาจำนวนหนึ่งสมมติว่า m ตัว ทำเครื่องหมาย และปล่อยปลาลงไปตามเดิม

– จับปลามาอีกโดยจะต้องจับให้ได้ปลาที่มีเครื่องหมายตามจำนวนที่ต้องการสมมติว่า k ตัว

ถ้าให้ X เป็นจำนวนปลาที่จับได้จนกว่าจะได้ปลาที่มีเครื่องหมาย k ตัว แล้ว X จะเป็นตัวแปรเชิงสุ่มไฮเปอร์จีโอเมทริคนิเสธที่มีการแจกแจงดังนี้

$$f(x) = \frac{\binom{m}{k-1}\binom{N-m}{x-k}}{\binom{N}{x-1}} \cdot \frac{m-k+1}{N-x+1} \; ; \quad x = k, k+1, \ldots$$

ในเมื่อ N เป็นจำนวนปลาทั้งหมดที่ไม่ทราบค่า

สำหรับการประมาณค่าของ N นั้นจะทำได้เช่นเดียวกับวิธีแรก ซึ่งจะได้ $\hat{N}$ ดังนี้

$$\hat{N} = [mx/k]$$

ตัวอย่าง จับปลาครั้งแรกมา 250 ตัว ทำเครื่องหมาย และปล่อยปลาลงไป ครั้งที่สองจับปลาจนได้ปลาที่มีเครื่องหมาย 50 ตัว ปรากฏต้องจับปลาทั้งหมด 5000 ตัว แล้วค่าประมาณของ N จะเป็น

$\hat{N} = [mx/_h[= [250(5000)/_{50}]$

$= 25000$ ตัว

ก.8 **การแจกแจงปัวซอง (Poisson distribution)** ในการทดลองเชิงสุ่มที่สนใจจำนวนครั้งที่เกิดขึ้นของเหตุการณ์หนึ่งในช่วงหรือเขต (interval or region) ที่กำหนดให้ แล้วตัวแปรเชิงสุ่ม X ที่แทนจำนวนครั้งที่เกิดขึ้นนี้จะมีการแจกแจงแบบนิวซอง ดังนี้

$$f(X) = \frac{\ell^{-\alpha} \cdot \alpha^{x}}{x!} \; ; \quad x = 0, 1, 2, \ldots$$

ในเมื่อ α เป็นส่วนเฉลี่ยของจำนวนครั้งที่เกิดขึ้นของเหตุการณ์ที่สนใจภายในช่วงที่กำหนดให้

ℓ เป็นค่าคงที่ และมีค่าประมาณ 2.71828......

ตัวแปรเชิงสุ่มพัวซองนี้เกิดขึ้นแพร่หลายในธรรมชาติ เช่น จำนวนอุบัติเหตุที่เกิดขึ้นในคาบเวลาหนึ่ง จำนวนคำที่พิมพ์ผิดต่อหน้าของหนังสือเล่มหนึ่ง จำนวนยานพาหนะที่เข้าเทียบเท่าเรือง ท่ารถยนต์ ท่าอากาศยาน เป็นต้น

ค่าคาดหวังและความแปรปรวนของตัวแปรเชิงสุ่มปัวซองนี้จะมีค่าเท่ากัน นั่นคือ

$E(X) = \alpha$

$V(X) = \alpha$

กราฟของการแจกแจงนี้ขึ้นอยู่กับ α นั่นคือจะเย้ทางขวาถ้า α มีค่าน้อย และจะสมมาตรถ้า α มีค่ามาก

ตัวอย่าง ในการศึกษาถึงอุบัติเหตุที่เกิดขึ้นกับคนงานในโรงงานอุตสาหกรรมอย่างหนึ่ง ปรากฏเฉลี่ยแล้วอาทิตย์ละ 0.8 ราย ถ้าจำนวนอุบัติเหตุต่ออาทิตย์เป็นตัวแปรเชิงสุ่มปัวซองแล้วจะมีการแจกแจงดังนี้

$$f(X) = \frac{\ell^{-0.8} \; (0.8)^{x}}{x!} \; ; \quad x = 0, 1, 2, \ldots$$

โอกาสที่อาทิตย์ใดอาทิตย์หนึ่งไม่มีอุบัติเหตุเลยจะเท่ากับ $f(0)$

$$= \ell^{-0.8}(0.8)^{0}/(0!) \quad = \ell^{-0.8}$$

$$= .44933$$

ในการแจกแจงทวินามนั้น ถ้าทำการทดลองมากครั้ง (n โต) เราจะคำนวนหาความน่าจะเป็นได้ยาก นักคณิตศาสตร์จึงหาวิธีอื่นช่วย วิธีหนึ่งที่ใช้กันก็คือใช้การแจกแจงปัวซอง ซึ่งสมัยแรกรู้จักในนามของ "สูตรคำนวนความน่าจะเป็นที่จะประสบความสำเร็จ x ครั้ง เมื่อทำ

การทดลองแบบเบอรนุยยี (Bernoulli's Trial) ซ้ำ ๆ กันหลาย ๆ ครั้ง" Simeon Denis Poisson (1781 — 1840) ได้พิสูจน์ว่า

$$\lim_{\substack{n \to \alpha \\ p \to 0}} \binom{n}{x} p^x q^{n-x} = \frac{e^{-np}(np)^x}{x!} = \frac{e^{-\alpha}\alpha^x}{x!}$$

ดังนั้นการแจกแจงปัวซอง นอกจากจะใช้อธิบายปรากฏการณ์ธรรมชาติแล้วมันยังใช้เป็นแบบขีดจำกัด (Limitting form) ของการแจกแจงทวินามอีกด้วย (ในเมื่อ n โต และ p น้อย)

โดยทั่วไปการประมาณการแจกแจงทวินามด้วยการแจกแจงปัวซองนั้นจะทำได้ถ้า $P \leqslant .05$ และ $n \geqslant 100$ การประมาณจะเที่ยงตรงขึ้นถ้า p น้อยมาก ๆ

ตัวอย่าง จากการศึกษาถึงคนติดสุราพบว่า โดยเฉลี่ยชายไทย 1000 คน จะติดสุรา 1 คน ถ้าศึกษาคน 8000 คน โอกาสที่จะพบคนติดสุราน้อยกว่า 7 คน เป็นเท่าใด?

ให้ X เป็นจำนวนคนที่ติดสุรา แล้ว X จะมีการแจกแจงดังนี้

$$f(x) = \binom{8000}{x} (.001)^x (.999)^{8000-x}$$

$$\approx e^{-8}\, 8^x / x! \quad ; \; \alpha = 8000\left(\frac{1}{1000}\right) = 8$$

ดังนั้น

$$P(X < 7) = \sum_{x=0}^{6} e^{-8}\, 8^x / x! = 0.3134$$

ข. กรณีตัวแปรเชิงสุ่มแบบต่อเนื่อง

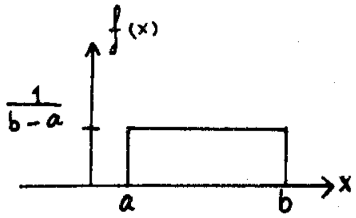
ข.1 การแจกแจงยูนิฟอร์ (Uniform distribution) ตัวแปรเชิงสุ่ม X ที่กำหนดจากการทดลองเชิงสุ่มนั้นมีค่าอยู่ในช่วงของเลขจำนวนจริง (a, b), a b และถ้าฟังก์ชั่นหนาแน่นน่าจะเป็นของ X นี้คงที่ตลอดช่วง (a, b) แล้วตัวแปรเชิงสุ่ม X จะมีการแจกแจงแบบยูนิฟอร์ม นั่นคือ

$$f(x) = \frac{1}{b-a} \quad ; \quad a < x < b$$

ค่าคาดหวังและความแปรปรวนจะเป็นดังนี้

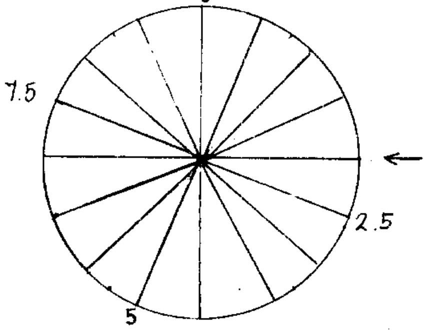
$E(X) = (b + a)/2$

$V(X) = (b - a)^2/12$



กราฟของการแจกแจง

ตัวอย่าง สมมติว่าหมุนวงล้อที่กำหนดให้ตามรูปข้างล่างนี้ และจะได้รับเงินตามที่ลูกศรชี้ ลูกศรนี้สามารถชี้ตรงเลขอะไรก็ได้ระหว่าง 0 ถึง 10 ถึงแม้ว่าในภาพจะแสดงแต่เลข 0, 2.5, 5, และ 7.5 เท่านั้น เช่น ถ้าลูกศรชี้กึ่งกลางระหว่าง 0 กับ 2.5 แล้วจะหมายถึงว่าลูกศรชี้ 1.25 เป็นต้น



ถ้าให้ X แทนจำนวนเงินที่ได้รับ แล้ว X จะเป็นตัวแปรเชิงสุ่มแบบยูนิฟอร์มที่มีการแจกแจงดังนี้

$$f(x) = \frac{1}{10}, \qquad 0 < x < 10$$

ดังนั้นโอกาสที่จะได้เงินระหว่าง 5 ถึง 8 จะเป็น $P(5 < X < 8)$

$$= \int_5^8 \frac{1}{10}\,dx \qquad = \left.\frac{x}{10}\right|_5^8 \qquad = .3$$

ค่าคาดหวังและความแปรปรวนของจำนวนเงินที่ได้รับคือ

$$E(X) = (b + a)/2 = (10 + 0)/2 = 5$$

$$V(X) = (b - a)^2/12 = (10 - 0)^2/12 = 100/12$$

$$= 8\tfrac{1}{3}$$

ข.2 การแจกแจงแบบปกติ (Normal or Gaussian distribution) การแจกแจงปกตินี้เป็นการแจกแจงที่มีความสำคัญมากเพราะการแจกแจงแบบต่าง ๆ ของตัวแปรเชิงสุ่ม สามารถประมาณได้ด้วยการแจกแจงแบบปกติโดยอาศัยตัวแปลง (Transformations) ง่าย ๆ และโดยเฉพาะในทางการศึกษาและจิตวิทยาได้นำการแจกแจงนี้ไปใช้กันมาก .

ตัวแปรเชิงสุ่ม X จะเรียกว่าแจกแจงแบบปกติ ถ้าฟังก์ชั่นหนาแน่นน่าจะเป็นของมันกำหนดไว้ดังนี้

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; \; -\infty < x < \infty$$

ในเมื่อ μ, σ เป็นพารามิเตอร์ที่มีค่าดังนี้ $\sigma > 0$, $-\infty \mu < \infty$ ส่วน π เป็นค่าที่มีค่าเท่ากับ 3.14159........

สำหรับค่าคาดหวังและความแปรปรวนของตัวแปรเชิงสุ่มปกติ คือ

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

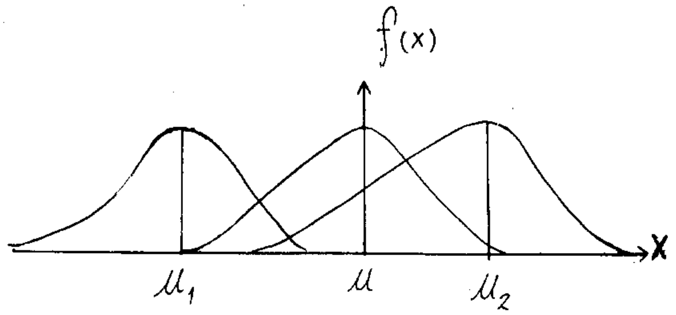
การแจกแจงปกติมีคุณลักษณะที่น่าสนใจดังนี้

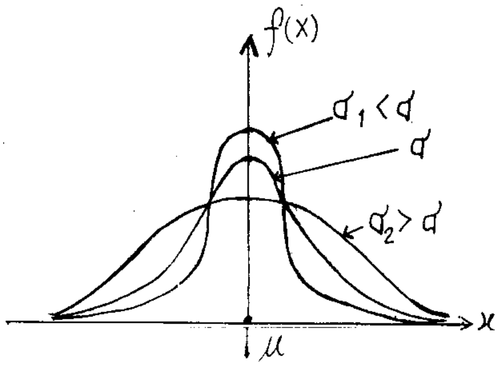
(1) จะสมมาตรที่ค่าคาดหวัง หรือ μ นั่นคือ $f(\mu+x) = f(\mu-x)$ กราฟของการแจกแจงปกติจะเป็นรูประฆังซึ่ง

(2) กราฟจะสูงสุดเมื่อ $x = \mu$ จะลดลงเรื่อย ๆ เมื่อ x ห่างจาก μ เพิ่มขึ้น ๆ ดังนั้น μ ซึ่งเป็นค่าคาดหวังก็จะเป็นฐานนิยมและมัธยฐานอีกด้วย

(3) รูปร่างของกราฟจะขึ้นอยู่กับ σ ถ้า σ ลดลงส่วนโค้งจะแคบเข้า นั่นคือถ้ากำหนดค่าคงที่ $k > 0$ ให้แล้วพื้นที่ใต้โค้งซึ่งอยู่ระหว่าง $\mu - k$ กับ $\mu + k$ จะเพิ่มขึ้นถ้า σ ลดลง

ส่วนการเปลี่ยนแปลงของ μ จะไม่ทำให้รูปร่างของกราฟเปลี่ยนแปลง เพียงแต่ย้ายโค้งไปทางขวาหรือซ้ายเท่านั้น คือถ้า μ เพิ่มขึ้น จะย้ายไปทางขวา และถ้า μ ลดลงก็จะย้ายไปทางซ้าย





(4) พื้นที่ใต้โค้งระหว่าง $\mu \pm \sigma$ เท่ากับ .6826, $\mu \pm 2\sigma$ เท่ากับ .9545, และ $\mu \pm 3\sigma$ เท่ากับ .9973

ในทางปฏิบัติที่ใช้กันบ่อย ๆ คือพื้นที่ใต้โค้งระหว่าง $\mu \pm 1.96\ \sigma$ เท่ากับ .95 และ $\mu \pm 2.85\ \sigma$ เท่ากับ .99

ในการคำนวนความน่าจะเป็นที่ตัวแปรเชิงสุ่มปกติ X จะมีค่าอยู่ในช่วง [a, b] เราก็หาได้ดังนี้

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \text{ สำหรับ } \mu$$

และ σ ที่กำหนดให้

เราจะเห็นได้ว่ายากที่จะคำนวน แต่เราก็มีวิธีง่าย ๆ ที่จะคำนวนความน่าจะเป็นของช่วงที่ต้องการนั้นได้โดยการใช้ตารางพิเศษจากการแจกแจงที่ได้ชื่อว่าการแจกแจงปกติมาตรฐาน

(Standard Normal distribution) ซึ่งดัดแปลงมาจากการแจกแจงปกติธรรมดานั่นเอง ตัวแปลงนั้นกำหนดไว้ดังนี้

$$Z = \frac{x-\mu}{\sigma}$$

ในเมื่อ Z เป็นตัวแปรเชิงสุ่มใหม่ที่ขึ้นอยู่กับตัวแปรเชิงสุ่มปกติ X เราสามารถจะบอกได้ว่า Z มีการแจกแจงปกติมาตรฐาน ดังทฤษฎีต่อไปนี้

ทฤษฎี ถ้าตัวแปรเชิงสุ่ม X มีการแจกแจงปกติ แล้วตัวแปรเชิงสุ่ม $Z = \frac{x-\mu}{\sigma}$ จะมีการแจกแจงปกติมาตรฐานที่มีค่าคาดหวังและความแปรปรวนเป็น 0 และ 1 ตามลำดับ และจะมีฟังก์ชั่นหนาแน่นน่าจะเป็นดังนี้

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-Z^2/2} \quad ; -\infty < Z < \infty$$

จากฟังก์ชั่นหนาแน่นปกติมาตรฐานนี้จะเห็นได้ว่าไม่มีพารามิเตอร์อยู่เลย ดังนั้นกราฟของมันจึงคงที่ และตารางที่สร้างขึ้นจากฟังก์ชั่นนี้จะคงที่ด้วย ซึ่งเป็นประโยชน์ในการประมาณค่าความน่าจะเป็นของช่วงที่ต้องการจากการแจกแจงปกติธรรมดาได้ด้วย แต่ต้องแปลงการแจกแจงปกติธรรมดาให้เป็นแบบมาตรฐานเสียก่อน ตารางนั้นมักเรียกว่า **ตารางปกติ หรือพื้นที่ใต้โค้งปกติ**

ตัวอย่าง ผลการสอบไล่ของนักเรียน ป. 2 จำนวน 500 คน ปรากฏว่ามีค่าเฉลี่ย 200 และส่วนเบี่ยงเบนมาตรฐาน 30 คะแนน

(1) จงหาเปอร์เซ็นต์ที่นักเรียนได้คะแนนมากกว่า 225 ระหว่าง 150 กับ 200

(2) ถ้านักเรียนที่ได้คะแนนตั้งแต่ 250 คะแนนขึ้นไปถือว่าเป็นนักเรียนระดับดีจะมีสิทธิได้รับรางวัล แล้วจะมีนักเรียนระดับดีกี่คน?

(3) ถ้าแบ่งนักเรียนเป็น 3 กลุ่ม คือกลุ่มเรียนดี, ปานกลาง และต้องแก้ไข โดยมีเปอร์เซ็นต์ดังนี้ 40, 30, 50 และ 20 ตามลำดับ แล้วกลุ่มเรียนปานกลาง V จะต้องได้คะแนนอย่างน้อยเท่าใด?

กลุ่มเรียนดี

ถ้าให้ X เป็นตัวแปรเชิงสุ่มปกติที่แทนคะแนนสอบไล่ และให้ $Z = \frac{x-\mu}{\sigma}$ เป็นตัวแปรเชิงสุ่มปกติมาตรฐาน แล้วเราจะแก้ปัญหาต่าง ๆ ได้ดังนี้

(1) $P(X > 225) = P\left(Z > \frac{225-200}{30}\right) = P(Z > 0.83)$

$= .5 - .2967 = .2033$

$$P(150 < X < 200) = P\left(\frac{150-200}{30} < Z < \frac{200-200}{30}\right)$$

$$= P(-1.67 < Z < 0) = .4525$$

(2) $$P(X \geq 250) = P\left(Z \geq \frac{250-200}{30}\right) = P(Z \geq 1.67)$$

$$= .5 - .4525 = .0475$$

ดังนั้นจะมีนักเรียนระดับดีเท่ากับ 500(.0475) $\cong$ 24 คน

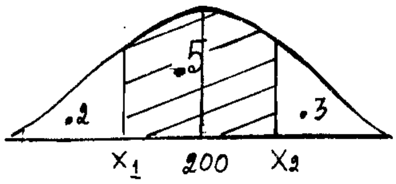
(3) ถ้าให้ X_1 และ X_2 เป็นคะแนนน้อยสุดที่กลุ่มเรียนปานกลาง และเรียนดีได้ แล้วเราจะได้ว่า (ดูรูป)

$$P(X < x_1) = .20$$

และ $$P(X \geq x_2) = .30$$

หรือ $$P\left(Z < \frac{x_1 - 200}{30}\right) = .20$$

และ $$P\left(Z \geq \frac{x_2 - 200}{30}\right) = .30$$



จากตารางเราได้ว่า $P(Z < .84) = .20$

และ $P(Z \geq .525) = .30$

ดังนั้น $\frac{x_1 - 200}{30} = -.84$ และ $\frac{x_2 - 200}{30} = .525$

นั่นคือ $x_1 = 174.8$ และ $x_2 = 215.75$

การแจกแจงปกตินั้นได้พัฒนามาเป็นระยะดังนี้ ในปี 1733 Abra ham de Moivre ได้นำเอาการแจกแจงปกติมาใช้ประมาณการแจกแจงทวินามที่มี p = 1/2 และ n มีค่าโต ในปี 1774 Pierre S. Laplace (1749–1827) ได้พิสูจน์ว่าการแจกแจงปกติสามารถใช้ประมาณความน่าจะเป็นของการแจกแจงทวินามที่มีค่า p ต่าง ๆ และ n มีค่าใหญ่ และยังใช้ประมาณความน่าจะเป็นของการแจกแจงปัวซองอีก ต่อมาปี 1809 Karl F. Gauss (1777–1855) พบว่าจำนวนเลขหลาย ๆ จำนวนซึ่งได้มาจากการวัด เช่นความสูงของคนหลาย ๆ คน จำนวนเหล่านี้เมื่อนำมาจัดเข้าในตารางแจกแจงความถี่ กราฟของการแจกแจงมักจะอยู่ในรูปของสมการโค้งปกติเสมอ ในระยะแรก ๆ การแจกแจงแบบนี้มักเรียกกันว่าการแจกแจงแบบเกาส์ (Gaussian distribution) การแจกแจงนี้เองที่ Gauss ได้นำมาใช้เป็นรากฐานของทฤษฎีว่าด้วยความเคลื่อนคลาด (Theory of Error) ของการวัดหรือการนับจำนวนใด ๆ

ทฤษฎีที่สำคัญที่ทำให้การแจกแจงปกติมีความสำคัญตามที่กล่าวมาแล้วในตอนแรก คือ "ทฤษฎีว่าด้วยขีดจำกัดสู่ส่วนกลาง (Central Limit Theorem)" ซึ่งกล่าวไว้คร่าว ๆ ว่าฟังก์ชั่น

แจกแจงน่าจะเป็นใด ๆ สามารถประมาณได้ด้วยฟังก์ชั่นแบบปกติเสมอ (All distribution converges to normal distribution) ในเมื่อมีเงื่อนไขบางประการในขีดจำกัด

การประมาณความน่าจะเป็นจากการแจกแจงทวินามและปัวซองนั้นเป็นกรณีพิเศษของทฤษฎีข้างบนนี้ การประยุกต์ที่สำคัญที่สุดของการแจกแจงปกติและทฤษฎีขีดจำกัดสู่ส่วนกลางก็คือการประยุกต์ในการสุ่มตัวอย่าง ซึ่งทฤษฎีขีดจำกัดเข้าสู่ส่วนกลางยืนยันว่า "การแจกแจงของส่วนเฉลี่ยจากตัวอย่างจะมีลักษณะเป็นโค้งปกติเสมอ หากขนาดของตัวอย่างใหญ่พอ"

ดังนั้นจึงขอกล่าวทฤษฎีขีดจำกัดสู่ส่วนกลางในรูปแบบสมบูรณ์ดังนี้

ทฤษฎีขีดจำกัดสู่ส่วนกลาง (Central Limit theorem) ถ้า $X_1, X_2, \ldots\ldots, X_n$ เป็นตัวแปรเชิงสุ่มที่เป็นอิสระกันและต่างก็มีการแจกแจงเหมือนกัน โดยมีค่าคาดหวังและความแปรปรวนเหมือนกัน นั่นคือ

$$E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$$

และ $$V(X_1) = V(X_2) = \ldots = V(X_n) = \sigma^2$$

แล้ว

$$\lim_{n \to \infty} P\left\{a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

$$= P(a \leq Z \leq b)$$

ในเมื่อ $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ เป็นฟังก์ชั่นหนาแน่นของตัวแปรเชิงปกติมาตรฐาน Z

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n); \quad E(\bar{X}) = \mu; \; V(\bar{X}) = \sigma^2/n$$

หรือกล่าวได้ว่า "การแจกแจงของ $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ จะเป็นการแจกแจงปกติมาตรฐาน ถ้า n โตพอ"

ต่อไปเราจะใช้การแจกแจงปกติไปประมาณการแจกแจงทวินาม และปัวซอง

(1) การประมาณการแจกแจงทวินาม ได้กล่าวมาแล้วว่าการแจกแจงปกติเป็นรูปฟอร์มขีดจำกัดของการแจกแจงหลายอย่าง ในกรณีของการแจกแจงทวินามที่มี n ขนาดใหญ่ แสดงได้ดังนี้

$$\lim_{n \to \infty} \binom{n}{x} p^x q^{n-x} \simeq \frac{1}{\sqrt{npq}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-np}{\sqrt{npq}}\right)^2}$$

หรือ

$$\lim_{n \to \infty} P\left(a < \frac{x - np}{\sqrt{npq}} < b\right) \cong P(a \leq Z \leq b)$$

ในเมื่อ Z เป็นตัวแปรเชิงสุ่มปกติมาตรฐาน

การประมาณนี้จะถูกต้องดี ถ้า $n \rightarrow \infty$ และ $p \rightarrow 1/2$ ถึงแม้ว่า n เล็ก และ p ไม่ใกล้ 0 หรือ 1 เกินไป ก็ยังใช้ได้อยู่ แต่จะให้ได้ผลเป็นที่น่าพอใจแล้ว np หรือ nq ควรจะมีค่ามากกว่า 5

G.W.Cochran ไม่ให้ข้อแนะนำในการประมาณแจกแจงทวินาม ด้วยการแจกแจงปกติไว้ดังนี้

ค่าของ p	ขนาดของ n (อย่างน้อย)
.5	30
.4 หรือ .6	50
.3 หรือ .7	80
.2 หรือ .8	200
.1 หรือ .9	600
.05 หรือ .95	1400

ขนาดของ n จะน้อยกว่าก็ได้ แต่การปรกมาณจะมีข้อผิดพลาดเพิ่มขึ้น

ในการประมาณการแจกแจงทวินามด้วยการแจกแจงปกตินั้นเป็นการประมาณการแจกแจงของตัวแปรเชิงสุ่มไม่ต่อเนื่องด้วยตัวแปรเชิงสุ่มต่อเนื่อง โดยที่ $P(X = a) = 0$ สำหรับตัวแปรเชิงสุ่มต่อเนื่อง X แต่ $P(X = a)$ อาจจะมากกว่า 0 ถ้า X ไม่ต่อเนื่อง ดังนั้นจึงจำเป็นต้องแก้ไขจุดปลาย (End points) การแก้ไขเช่นนี้เรียกว่า **การแก้ไขเพื่อการต่อเนื่อง** (Correction for continuity) ซึ่งทำได้ดังนี้

ก. $$P(X = a) \cong P\left(\frac{a - .5 - np}{\sqrt{npq}} < z < \frac{a + .5 - np}{\sqrt{npq}}\right)$$

ข. $$P(a \leqslant X \leqslant b) \cong P\left(\frac{a - .5 - np}{\sqrt{npq}} < z < \frac{b + .5 - np}{\sqrt{npq}}\right)$$

ตัวอย่าง เป็นที่ทราบกันว่าแม่บ้าน 10% จะซื้อหวยเถื่อน ถ้าในวันหวยเถื่อนออกสุ่มแม่บ้านมา 18 ราย แล้วโอกาสที่จะมีแม่บ้าน 3 ถึง 5 ราย จะซื้อหวยเถื่อน เป็นเท่าใด?

เมื่อใช้การแจกแจงทวินามเราได้ $$P(3 \leqslant X \leqslant 5) = \sum_{x=3}^{5} \binom{18}{x} (.1)^x (.9)^{18-x}$$

$$= f(3) + f(4) + f(5)$$

$$= 0.168 + 0.07 + 0.0218 = 0.2598$$

เมื่อประมาณด้วยโค้งปกติ

$$P(3 \leqslant X \leqslant 5) = P\left(\frac{3-.5-1.8}{\sqrt{18(.1)(.9)}} < Z < \frac{5+.5-1.8}{\sqrt{18(.1)(.9)}}\right)$$

$$= P(.55 < Z < 2.91)$$

$$= .4982 - .2088 = .2894$$

(2) การประมาณการแจกแจงปัวซอง โดยที่การแจกแจงปัวซองใช้ประมาณการแจกแจงทวินามเมื่อ n โต (np ไม่น้อยเกินไป) ดังนั้นจึงกลายเป็นว่าการแจกแจงปกติจะประมาณการแจกแจงปัวซองเมื่อค่าเฉลี่ย α โตพอ และ Laplace ก็ได้พิสูจน์ไว้ว่า

$$\lim_{\alpha \to \infty} e^{-\alpha}\alpha^{x}/x! \cong \frac{1}{\sqrt{\alpha}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\alpha}{\sqrt{\alpha}}\right)^{2}}$$

หรือ

$$\lim_{\alpha \to \infty} P(a \leqslant x \leqslant b) = P\left(\frac{a-.5-\alpha}{\sqrt{\alpha}} < Z < \frac{b+.5-\alpha}{\sqrt{\alpha}}\right)$$

ตัวอย่าง สมมติว่าจำนวนนักศึกษาที่ใช้เวลาสอบในวิชาสถิติไม่เกิน 1 ชั่วโมง มีการแจกแจงปัวซอง ที่มี $\alpha = 10$ แล้วโอกาสที่นักศึกษา 9 ถึง 12 คนจะใช้เวลาสอบไม่เกิน 1 ชั่วโมงเป็นเท่าใด?

ใช้การแจกแจงปัวซอง $P(9 \leqslant X \leqslant 12) = \sum_{x=9}^{12} e^{-10}\, 10^{x}/x!$

$$= .459$$

ใช้การแจกแจงปกติ $P(9 \leqslant X \leqslant 12) = P\left(\frac{9-.5-10}{\sqrt{10}} < Z < \frac{12+.5-10}{\sqrt{10}}\right)$

$$= .360$$

ข.3 การแจกแจงเอกซ์โพเน็นท์ (Exponential distribution) ตัวแปรเชิงสุ่ม X ซึ่งเป็นคาบเวลาของการรอคอยการเกิดขึ้นครั้งแรกหรือสองครั้งติดต่อกันของเหตุการณ์ที่สนใจในช่วงเวลาที่กำหนดนั้นจะมีการแจกแจงเอกซ์โพเน็นท์ ดังนี้

$$f(x) = \alpha e^{-\alpha x} \;; \quad x > 0$$

ฟังก์ชั่นแจกแจงค่าคาดหวังและความแปรปรวนกำหนดไว้ดังนี้

$$F(X) = 1 - e^{-\alpha x} \;; \quad x > 0$$

$$E(X) = 1/\alpha$$

$$V(X) = 1/\alpha^{2}$$

ตัวแปรเชิงสุ่มเอกซ์โพเน็นท์กับปัวซองมีความสัมพันธ์กัน ดังนี้ ถ้า Y เป็นตัวแปรเชิงสุ่มปัวซองที่เป็นจำนวนครั้งของการเกิดขึ้นของเหตุการณ์หนึ่งในคาบเวลาที่กำหนด แล้วตัวแปรเชิงสุ่มเอกซ์โพเน็นท์ X จะเป็นคาบเวลาของการรอคอยการเกิดขึ้นครั้งแรกหรือสองครั้งติดต่อกันของเหตุการณ์นั้น

การแจกแจงเอกซ์โพเน็นท์มีบทบาทสำหคัญในปัญหาการรอคอย (Queueing Problems) และทฤษฎีเชื่อมั่น (Reliability Theory)

ตัวอย่าง ให้ Y เป็นตัวแปรเชิงสุ่มที่แทนจำนวนครั้งที่ต่อโทรศัพท์ไปติดและมีการแจกแจงปัวซอง ถ้าเฉลี่ยแล้วต่อไม่ติด 2 ครั้งในหนึ่งชั่วโมง เราจะได้

$$f(y) = e^{-2}2^y/y! \; ; \; y = 0,1,2,\ldots$$

และให้ X เป็นเวลาระหว่างการต่อไม่ติด 2 ครั้งติดต่อกัน แล้วเราจะได้

$$f(x) = 2e^{-2x} \; ; \; x > 0$$

สมมติว่าโทรศัพท์เพิ่งต่อไม่ติด แล้วความน่าจะเป็นที่โทรศัพท์ต่อไม่ติดจะเกิดขึ้นอีกภายใน 2 ชั่วโมง หน้านี้จะเท่ากับ

$$P(0 \leq X \leq 2) = \int_0^2 2e^{-2x}\,dx = 1-e^{-4}$$

$$= .982$$

ข.4 การแจกแจงแกมมะ (Gamma distribution) การแจกแจงนี้เป็นแบบทั่วไปของการแจกแจงเอกซ์โพเน็นท์ นั่นคือถ้าเราทราบว่าตัวแปรเชิง X ที่แทนเวลาของการรอคอยเหตุการณ์จะเกิดขึ้นครั้งแรก แล้ว X มีการแจกแจงเอกซ์โพเน็นท์ สมมติถ้าเรากำหนดตัวแปรเชิงสุ่ม Y ซึ่งแทนเวลาทั้งหมดที่รอคอยเหตุการณ์จะเกิดขึ้น k ครั้ง แล้วสามารถแสดงได้ว่า

$$Y = X_1 + X_2 + \ldots + X_k$$

ในเมื่อ X_i เป็นเวลาที่รอคอยของการเกิดขึ้นครั้งที่ i ($i = 1, 2, \ldots, n$) โดยที่ X_i เป็นอิสระต่อกัน และฟังก์ชันหนาแน่นน่าจะเป็นของ Y จะเป็นดังนี้

$$f(y) = \frac{\alpha}{\Gamma(k)}(\alpha y)^{n-1}e^{-\alpha x} \; ; \; y \geq 0,\ \alpha > 0,\ n \geq 1$$

นั่นคือเวลาทั้งหมด Y ของการเกิดขึ้น n ครั้งติดต่อกันจะมีฟังก์ชันหนาแน่นน่าจะเป็นแบบแกมมะ $f(y)$ โดยกำหนดว่าเวลาที่รอคอยภายในการเกิดขึ้น 2 ครั้งติดต่อกันเป็นแบบเอกซ์โพเน็นท์

สำหรับ $\Gamma(k)$ นั้นเรียกว่าฟังก์ชันแกมมะ (Gamma function) ซึ่งกำหนดไว้ดังนี้

$$\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x}\, dx$$
$$= (k-1)\,\Gamma(k-1) = (k-1)!$$

และ $\Gamma(1/2) = \sqrt{\pi}$

ค่าคาดหวังและความแปรปรวนของการแจกแจงแกมมะจะเป็น

$$E(y) = n/\alpha$$
$$V(y) = n/\alpha^2$$

ถ้า $k = 1$, $f(y) = \alpha e^{-\alpha y}$ ซึ่งเป็นฟังก์ชั่นหนาแน่นเอกซ์โพเน็นท์

ถ้า $\alpha = 1/2$, $n = \nu/2$ และ ν เป็นค่าเต็มบวก แล้วตัวแปรเชิงสุ่ม Y จะมีการแจกแจงไคสแควร์ (Chi-Square) ที่มีองศาแห่งความเป็นอิสระ ν นั่นคือ

$$f(y) = \frac{1/2}{\Gamma(\nu/2)} (y/2)^{\nu/2-1} e^{-y/2}$$

การแจกแจงแกมมะบางครั้งเรียกว่า Erlangian or Pearson's Type III distribution

นอกจากการแจกแจงแกมมะมีความสัมพันธ์กับการแจกแจงไคสแควร์ที่มีบทบาทสำคัญในทฤษฎีสถิติ แล้วการแจกแจงแกมมะยังมีการประยุกต์ในทฤษฎีสถิตินวสมัย (Modern Statistics or Bayesiam Statistics) และปัญหาการตัดสินใจ

ข.5 การแจกแจงบีทะ (Beta distribution) ตัวแปรเชิงสุ่มบางตัว (เช่นเปอร์เซ็นต์หรือสัดส่วนของเหตุการณ์หนึ่ง) อาจจะมีค่าอยู่ในช่วง (1,/1) และความน่าจะเป็นของช่วงย่อย (Sub-interval) ที่เท่ากันในช่วง 0, 1) จะไม่เท่ากัน (ถ้าเท่ากันแล้วตัวแปรเชิงสุ่มนั้นจะเป็นแบบยูนิฟอร์ม) ตัวแปรเชิงสุ่มที่มีลักษณ์เช่นนี้จะมีการแจกแจงบีทะ อันมีฟังก์ชั่นหนาแน่นดังนี้

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}; \quad 0<x<1,\ \alpha,\beta>0$$

ในเมื่อ $B(\alpha, \beta)$ เรียกว่าฟังก์ชั่นบีทะ (Beta function) กำหนดไว้ดังนี้

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$$
$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

ค่าคาดหวังและความแปรปรวนจะเป็น

$$E(X) = \frac{\alpha}{\alpha+\beta}$$

$$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

ถ้า $\alpha = 1, \beta = 1$ การแจกแจงจะกลายเป็นการแจกแจงยูนิฟอร์ม

การแจกแจงบีทะนี้ในปัจจุบันใช้ประยุกต์ในทฤษฎีสถิตินวสมัย และทฤษฎีตัดสินใจเป็นอย่างมาก

ข.6 การแจกแจงพาเรโต (Pareto distribution) การแจกแจงนี้ตั้งตามชื่อของ Vilfredo Pareto (1848 — 1923) นักสังคมวิทยาและเศรษฐศาสตร์ชาวอิตาลี ซึ่งมีความสนใจในการแจกแจงรายได้ (Income distribution) และได้ตั้งกฎที่ชื่อว่า "Pareto's Law" ที่เกี่ยวกับการแจกแจงของรายได้ต่อประชากร (distribution of income over a population) ดังนี้

$$N = Ax^{-a}$$

ในเมื่อ N เป็นจำนวนบุคคลที่มีรายได้เท่ากับหรือมากกว่า X, A และ a เป็นพารามิเตอร์ โดยที่ a เรียกว่าค่าคงที่พาเรโต (Pareto's Constant)

การแจกแจงพาเรโตนี้ใช้อธิบายปรากฏการณ์ต่าง ๆ เช่น การแจกแจงระดับรายได้, ขนาดของเมือง, และความถี่ของคำ เป็นต้น การแจกแจงนี้กำหนดให้ฟังก์ชั่นหนาแน่นไว้ดังนี้

$$f(x) = (\alpha/x_0)(x_0/x)^{\alpha+1} \; ; \quad x > x_0 \; ; \; x_0, \alpha > 0$$

ค่าคาดหวัง, มัธยฐาน, และความแปรปรวนจะเป็น

$$E(X) = \frac{\alpha x_0}{\alpha - 1} \; ; \quad \alpha > 1$$

$$\mu_m = \sqrt[\alpha]{2}\; x_0$$

$$V(X) = \frac{\alpha x_0^2}{(\alpha-1)^2(\alpha-2)} \; ; \quad \alpha > 2$$

ข.7 การแจกแจงไคสแควร์ (Chi — Square, X^2, distribution) การแจกแจงนี้เป็นรูปหนึ่งของการแจกแจงแกมมะ นั่นคือฟังก์ชั่นหนาแน่นน่าจะเป็นแบบไคสแควร์จะเป็นดังนี้

$$f(x) = \frac{1/2}{\Gamma(\nu/2)}(x/2)^{\nu/2-1} e^{-x/2} \; ; \quad x > 0$$

ในเมื่อ ν เป็นค่าคงที่และเป็นจำนวนเต็มบวก ซึ่งเรียกว่าองศาแห่งความเป็นอิสระ(DEGREE OF FREE) การแจกแจงไคสแควร์กับการแจกแจงปกติมาตรฐานมีความสัมพันธ์กันดังนี้ "ถ้า

$Z_1, Z_2, \ldots, Z_n$ เป็นตัวแปรเชิงสุ่มปกติมาตรฐานที่เป็นอิสระกัน แล้ว $X = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ จะเป็นตัวแปรเชิงสุ่มที่มีการแจกแจงแบบไคสแควร์ที่มีองศาความเป็นอิสระเท่ากับ n"

การแจกแจงไคสแควร์นี้ F.R.Helmert เป็นผู้พัฒนา และ Karl Rearson เป็นผู้นำไปประยุกต์เกี่ยวกับการทดสอบสมมติฐาน ในสถิติอนุมานได้อาศัยการแจกแจงนี้อย่างมาก และตารางของการแจกแจงนี้ก็ใช้กันแพร่หลาย

สำหรับค่าคาดหวังและความแปรปรวนของการแจกแจง จะเป็น

$$E(X) = \nu$$

$$V(X) = 2\nu$$

ข.8 การแจกแจงแบบที่ (Student's t–distribution) ตัวแปรเชิงสุ่ม X จะเรียกว่ามีการแจกแจงแบบ ที ถ้าฟังก์ชั่นหนาแน่นกำหนดไว้เป็น

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\nu\pi}} \cdot \left(1 + x^2/\nu\right)^{-\frac{\nu+1}{2}}; \quad -\infty < x < \infty$$

ในเมื่อ ν เป็นองศาความเป็นอิสระ

ค่าคาดหวังและความแปรปรวนจะเป็น

$$E(x) = 0$$

$$V(x) = \frac{\nu}{\nu-2}, \quad \nu > 2$$

การแจกแจงแบบทีนี้มีความสัมพันธ์กับการแจกแจงไคสแควร์ และ การแจกแจงปกติมาตรฐานดังนี้ "ถ้าตัวแปรเชิงสุ่ม Z มีการแจกแจงปกติมาตรฐาน, ตัวแปรเชิงสุ่ม U มีการแจกแจงไคสแควร์ที่มีองศาความเป็นอิสระ ν และ Z กับ U เป็นอิสระกัน แล้วตัวแปรเชิงสุ่ม $T = Z/\sqrt{U/\nu}$ จะมีการแจกแจงแบบทีซึ่งมีองศาความเป็นอิสระ ν "

กราฟของการแจกแจงแบบทีจะเป็นรูประฆังคล้ายโค้งปกติมาตรฐาน แต่แบนกว่าเล็กน้อย แต่ถ้า ν ยิ่งโต กราฟก็จะใกล้โค้งปกติมาตรฐานยิ่งขึ้น นั่นคือ เมื่อ $\nu \to \infty$ การแจกแจงแบบทีจะกลายเป็นการแจกแจงแบบปกติมาตรฐาน

การแจกแจงแบบทีนี้มีบทบาทต่อสถิติอนุมานและการวิจัยเป็นอย่างมากตารางของการแจกแจงก็ใช้กันแพร่หลาย William Sealy Gosset เป็นผู้พัฒนาการแจกแจงแบบทีในปี 1908 ผล

งานนั้นชื่อว่า "The Probable Error of a Mean" ลงในวารสาร Biometrika โดยใช้นามปากกาว่า "Student"

ข.9 การแจกแจงแบบเอฟ (Snedecor's F—distribution) การแจกแจงนี้มีความสำคัญในสถิติการทดลองอย่างมาก G.W.Snedecor เป็นผู้พัฒนาขึ้นมา และตั้งชื่อให้เป็นเกียรติแก่ Ronald Fisher ว่าการแจกแจงแบบเอฟ เพราะ Fisher ได้เป็นผู้ศึกษาครั้งแรกในปี 1924

ถ้าตัวแปรเชิงสุ่ม U และ V เป็นอิสระกันและต่างก็มีการแจกแจงไคสแควร์ ซึ่งมีองศาความเป็นอิสระ ν_1 และ ν_2 ตามลำดับ แล้วตัวแปรเชิงสุ่ม $X = (U/\nu_1)/(V/\nu_2)$ จะเป็นตัวแปรเชิงสุ่มแบบเอฟที่มีองศาแห่งความเป็นอิสระ $(\nu_1\ \nu_2)$ และมีฟังก์ชันหนาแน่นดังนี้

$$f(X) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \left(\nu_1/\nu_2\right)^{\nu_1/2} X^{\left(\frac{\nu_1}{2}-1\right)} \left(1+\frac{\nu_1}{\nu_2}X\right)^{-\frac{\nu_1+\nu_2}{2}} ;\ X>0$$

การแจกแจงแบบเอฟมีคุณสมบัติที่น่าสนใจดังนี้

(1) ค่าคาดหวังและความแปรปรวนจะเป็น

$$E(X) = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2$$

$$V(X) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, \quad \nu_2 > 4$$

(2) กราฟของการแจงจะเย้ทางขวา ความเย้จะลดลง ถ้าจำนวนองศาความเป็นอิสระเพิ่มขึ้น

(3) การแจกแจงแบบเอฟนี่มีคุณสมบัติของส่วนกลับได้ นั่นคือ "ตัวแปรเชิงสุ่ม Y = 1/X จะมีการแจกแจงแบบเอฟที่มีองศาความเป็นอิสระ (ν_2, ν_1) ซึ่งกลับกันกับของ X ซึ่งมี (ν_1, ν_2)

(4) การแจกแจงแบบที กับแบบเอฟมีส่วนสัมพันธ์กันดังนี้ "ถ้า X มีการแจกแจงแบบทีด้วยองศาความเป็นอิสระ k แล้ว X^2 จะมีการแจกแจงแบบเอฟด้วยองศาความเป็นอิสระ (1, k) นั่นคือ

$$P(F^{(1,n)} > a^2) = P(t^{(n)} > a) + P(t^{(n)} < -a) = 2P(t^{(n)} > a)"$$

(5) การแจกแจงไคสแควร์กับแบบเอฟมีความสัมพันธ์กันดังนี้ "ถ้า y มีการแจกแจงแบบเอฟด้วยองศาความเป็นอิสระ ν_1, ∞ แล้ว $\nu_1 \times$ จะมีการแจกแจงไคสแควร์ด้วยองศาความเป็นอิสระ m นั่นคือ

$$P(F^{(\nu_1,\infty)} > a/m) = P(\chi^{2(\nu_1)} > a)"$$

(6) การแจกแจงทวินามกับแบบเอฟมีความสัมพันธ์กันดังนี้ "ให้ x เป็นตัวแปรเชิงสุ่มทวินามที่มีพารามิเตอร์ n และ P และให้ r เป็นค่าเลขเต็มบวกที่ทำให้

$$P(x \geqslant r) \leqslant \alpha < P(x \geqslant r-1)$$

แล้วเราจะได้ว่า

$$P(x \geqslant r) = P\left(y \geqslant \frac{r}{n+1-r} \; \frac{1-P}{P}\right)$$

ในเมื่อ y เป็นตัวแปรเชิงสุ่มแบบเอฟด้วยองศาความเป็นอิสระ 2(n + 1 – r) และ 2r

สำหรับ r เมื่อให้เป็นค่าที่ทำให้ $P(x \leqslant r) \leqslant \alpha < P(x \leqslant r+1)$ แล้วเราจะได้อีกว่า

$$P(x \leqslant r) = P\left(Z \geqslant \frac{n-r}{r+1} \; \frac{P}{1-P}\right)$$

ในเมื่อ Z เป็นตัวแปรเชิงสุ่มแบบเอฟด้วยองศาความเป็นอิสระ 2(r + 1) และ 2(n – r)

จากความสัมพันธ์นี้เราสามารถใช้ประมาณค่าและทดสอบสมมติฐานเกี่ยวกับค่า P นี้ได้

(11) **การแจกแจงปกติแบบล็อก** (Lognormal Distribution) ถ้าให้ x เป็นตัวแปรเชิงสุ่มที่เป็นบวก $(X > 0)$ และให้ y = ln x โดยที่ y มีการแจกแจงปกติ แล้ว x จะมีการแจกแจงปกติแบบล็อก ดังนี้

$$f(x) = (1/x\sigma\sqrt{2\pi})\, e^{-(1/2\sigma^2)(\ln x - \mu)^2}; \quad x>0; \mu>0; \sigma>0$$

ค่าคาดหวังและความแปรปรวน จะเป็นดังนี้

$$E(x) = e^{\mu + \sigma^2/2}$$

$$V(X) = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

$$E(\ln X) = \mu, \; V(\ln X) = \sigma^2$$

การแจกแจงแบบนี้จะมีฐานนิยมเดียว แต่เบ้ทางขวา เมื่อ σ^2 มีค่าน้อย แล้วรูปร่างของการแจกแจงจะใกล้กับโค้งปกติ

(12) **การแจกแจงปกติสองมิติ** (Bivariate Normal Distribution) ตัวแปรเชิงสุ่ม (X, Y) ซึ่งเป็นแบบต่อเนื่องจะมีการแจกแจงปกติสองมิติ ถ้าฟังก์ชั่นหนาแน่นน่าจะเป็นร่วมกำหนดไว้ดังนี้

$$f(x\,,y) = (1/2\pi\sigma_x\sigma_y\sqrt{1-\rho^2})\exp\left[-\frac{1}{2(1-\rho)^2}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\right]$$

การแจกแจงนี้มีค่าเฉลี่ย ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ ดังนี้

$$E(x) = \mu_n\,, E(y) = \mu$$

$$E(y/x) = \mu_y + \rho(\sigma_y/\sigma_x)(x-\mu_x)$$

$$V(x) = \sigma_x^2\,,\ V(y) = \sigma_y^2$$

$$\rho_{xy} = \rho\,,\ Cov(x,y) = \rho\sigma_x\sigma_y$$

เราจะเห็นได้ว่า E(y/x) เป็นฟังก์ชั่นเชิงเส้น (Linearfunction) ของ X