ปัญหาเกี่ยวกับการสุ่ม

พึงรีบเร่งกระทำความดี
และป้องกันจิตจากความชั่ว
เพราะเมื่อกระทำความดีช้าไป
ใจจะกลับยินดีในความชั่ว

พุทธวจนะในธรรมบท

บ่อยครั้งที่ต้องการทราบว่าอนุกรมของเหตุการณ์หรือสิ่งของนั้นเป็นแบบสุ่ม (Random) ข้อมูลตัวอย่างสำหรับการวิเคราะห์ทางสถิตินั้นก็เป็นแบบสุ่ม ข้อกำหนดเบื้องต้นภายใต้กระบวนการของสถิติอนุมานก็คือ การอ้างอิงที่ขึ้นอยู่กับตัวอย่างสุ่ม กระบวนการควบคุมคุณภาพ และการวิเคราะห์ถดถอยจะมีข้อกำหนดเกี่ยวกับการสุ่ม ถ้าสงสัยการสุ่มของตัวอย่าง ก็จำเป็นต้องมีวิธีที่จะตัดสินใจว่าตัวอย่างนั้นเป็นแบบสุ่มหรือไม่ก่อนที่จะดำเนินการวิเคราะห์ วิธีการหรือแบบทดสอบเกี่ยวกับการสุ่มจะมีดังต่อไปนี้

7.1 แบบทดสอบโดยอาศัยรัน (Run Test for Randomness)

รัน (Run) เป็นอนุกรมของสัญลักษณ์ที่เหมือนกันซึ่งอาจจะตามหรือนำสัญลักษณ์อื่นๆ หรือไม่มีสัญลักษณ์อื่นตามหรือนำเลยก็ได้ จำนวนสัญลักษณ์ในรันหนึ่งๆ ก็คือความยาวของรัน ถ้าปรากฏว่ามีจำนวนรันมากไปหรือน้อยไป แล้วอนุกรมจะไม่เป็นแบบสุ่ม ดังนั้นรันจึงใช้สำหรับทดสอบการสุ่มของการจัดเรียง และใช้ทดสอบสมมติฐานที่ว่าสองตัวอย่างสุ่มที่เป็นอิสระกันนั้นมาจากประชากรที่เป็นแบบเดียวกัน แบบทดสอบเกี่ยวกับการสุ่มโดยอาศัยรัน มีอยู่ 3 แบบ ดังนี้

(1) ผลรวมของรัน (Total number of Runs) ผลรวมของรัน R ใช้ทดสอบการสุ่มของอนุกรมสัญลักษณ์สองชนิด ค่าเฉลี่ย และความแปรปรวนของ R จะเป็น

$$E(R) = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$V(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

ในเมื่อ n_1 เป็นจำนวนสัญลักษณ์ประเภทหนึ่ง และ n_2 เป็นจำนวนสัญลักษณ์อีกประเภทหนึ่ง

เมื่อตัวอย่างขนาดโต (n_1 และ n_2 โตกว่า 20) การทดสอบการสุ่มจึงใช้ตัวสถิติซึ่งมีการแจกแจงปกติมาตรฐาน

$$Z = \frac{R - E(R)}{\sqrt{V(R)}}$$

ถ้าตัวอย่างขนาดเล็ก ก็จะอาศัยตารางพิเศษ (ตาราง 44) ในการกำหนดค่าวิกฤต

ตัวอย่าง ในการศึกษาเกี่ยวกับอันดับของเพศสำหรับลูกค้าที่เข้ามาติดต่อธนาคาร 50 ราย ได้ผล

ช ญ ช ญ ช   ช ช ญ ญ ช   ญ ช ญ ช ญ

ช ช ช ช ญ   ช ญ ช ญ ช   ช ญ ญ ญ ช

ญ ช ญ ช ญ   ช ช ญ ช ช   ญ ช ช ช ช

ญ ช ญ ช ช

อันดับเพศเป็นแบบสุ่มหรือไม่?

Ho : อันดับเพศของลูกค้าที่เข้ามาติดต่อธนาคารเป็นแบบสุ่ม

Ha : อันดับเพศไม่เป็นแบบสุ่ม

จากข้อมูลจะได้ $n_1 = 30$ $n_2 = 20$ และ $R = 35$

$$E(R) = 2(30)(20)/(30+20) + 1 = 25$$

$$V(R) = \frac{2(30)(20)\{2(30)(20)-30-20\}}{(30+20)^2(30+20-1)} = 11.265$$

$$Z = \frac{35-25}{\sqrt{11.265}} = 2.98$$

สำหรับ $Z_{.025} = 1.96$ จึงสรุปได้ว่า การเข้ามาติดต่อของลูกค้าไม่เป็นไปแบบสุ่ม

(2) รันที่อยู่เหนือและใต้มัธยฐาน (Runs above and below the Median) รันชนิดนี้ใช้ทดสอบการสุ่มของตัวอย่างโดยอาศัยอันดับที่ของค่าสังเกตจากตัวอย่างที่เลือกสุ่มมา ค่าแต่ละค่าจะแทนด้วยอักษร a หรือ b ตามแต่ว่ามันมีค่ามากกว่าหรือน้อยกว่ามัธยฐานของตัวอย่าง แล้วใช้การทดสอบของรันในแบบ (1) ประยุกต์เข้ากับอนุกรมของ a และ b

ถ้าจำนวนตัวอย่างเป็นจำนวนคู่และมากกว่า 25 โดยที่ประชากรเป็นแบบต่อเนื่อง แล้วการแจกแจงของจำนวนรัน R จะเป็นแบบปกติที่มีค่าเฉลี่ย และความแปรปรวน ดังนี้

$$E(R) = n/2 + 1$$

$$V(R) = n(n-2)/4(n-1)$$

นั่นคือตัวสถิติ Z จะมีการแจกแจงปกติมาตรฐาน

$$Z = \frac{R - E(R)}{\sqrt{V(R)}}$$

ตัวอย่าง จากการศึกษาผลการสอบสถิติของนักศึกษาโดยอาศัยตัวอย่างขนาด 26 ราย ปรากฏว่าได้คะแนนดังนี้

		97	89	25	81	11	83	16	96	44	32	98	19
68	33	25	54	74	82	17	49	33	22	62	20	92	80

มัธยฐานจะเป็น (54+49)/2 = 51.50 เมื่อแทนข้อมูลด้วย a หรือ b ก็จะได้ดังนี้

a a b a b a b a b b a b a b b d a a b b b b a b a a

$$n = 26 \quad n_1 = 13 \quad n_2 = 13 \quad R = 17$$

E(R) = 26/2 + 1 = 14

V(R) = 26(26-2)/4(26-l) = 6.24

$$Z = \frac{17 - 14}{\sqrt{6.24}} = 1.21$$

เนื่องจาก $Z_{.025}$ = 1.96 จึงสรุปได้ว่า ตัวอย่างนี้เป็นตัวอย่างสุ่ม

(3) **รันที่อยู่บนและล่าง (Runs up and down)** รันชนิดนี้ใช้ทดสอบการสุ่มเช่นเดียวกัน แต่พิจารณาค่าสังเกตทีละคู่ที่ติดต่อกัน แล้วแทนด้วยเครื่องหมาย + หรือ – ตามแต่ว่าค่าหลังมากกว่าหรือน้อยกว่าค่าแรก แล้วเราจะได้รัน เมื่อใช้การทดสอบรันในแบบ (1) เข้าประยุกต์กับอนุกรมของ + และ – ก็จะตัดสินใจได้ว่า ตัวอย่างเป็นแบบสุ่มหรือไม่

จำนวนรันของเครื่องหมาย + และ – นี้จะมีค่าเฉลี่ย และความแปรปรวน ดังนี้

$$E(R) = (1/3)(2n-1)$$

$$V(R) = (1/90)(16n-29)$$

เมื่อตัวอย่างขนาดโต ($n \geq 20$) แล้วจะได้ตัวสถิติ Z ซึ่งมีการแจกแจงปกติมาตรฐาน

$$Z = \frac{R - E(R)}{\sqrt{V(R)}}$$

ตัวอย่าง จากตัวอย่างที่ผ่านมา จะได้เครื่องหมาย + และ – ดังนี้

– – + – + – + – – + – + – – + + + – + – – – + – + –

ดังนั้น $n_1 = 14$ $n_2 = 11$ และ R = 19

E(R) = (1/3){2(26)–1} = 17

V(R) = (1/90){16(26)–29} = 4.3

$$Z = \frac{19 - 17}{\sqrt{4.3}} = 1.4$$

เมื่อ $\alpha = .05$ จะได้ $Z_{.025}$ = 1.96 จึงสรุปได้ว่า ตัวอย่างเป็นตัวอย่างสุ่ม

ฟามา (Eugene F. Fama, 1965) ได้เสนอตัวสถิติโดยอาศัยผลต่างระหว่างจำนวนรันจริงและที่คาดไว้ (Actual and Expected number of Runs) และรันนั้นเป็นรันที่เป็นบวก ลบ และศูนย์ ถ้าสมมติว่าสัดส่วนตัวอย่างของ บวก ลบ และศูนย์ เป็นตัวประมาณค่าที่ดีของสัดส่วนประชากร แล้วภายใต้สมมติฐานเกี่ยวกับความเป็นอิสระนั้นสามารถคำนวณจำนวนรันที่คาดไว้ของทุกเครื่องหมาย (+, –, 0) ได้เป็น

$$m = \{n(n+1) - \Sigma R_i^2\}/n$$

ในเมื่อ n เป็นจำนวนการเปลี่ยนแปลง (+, -, 0) หรือขนาดตัวอย่าง และ R_i เป็นจำนวนการเปลี่ยนแปลงของแต่ละเครื่องหมาย ความแปรปรวนของ m จะเป็น

$$V(m) = \frac{\Sigma R_i^2\{\Sigma R_i^2 - n(n+1) - 2n\Sigma R_i^3 - n^3}{n^2(n-1)}$$

เมื่อ n โตพอ แล้วการแจกแจงของ m จะเป็นแบบปกติโดยประมาณ

สำหรับตัวอย่างขนาดโต จำนวนรัน R จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น m และความแปรปรวนเป็น V(m) ดังนั้น Z จะมีการแจกแจงปกติมาตรฐาน ในเมื่อ

$$Z = \frac{(R - 0.5) - m}{\sqrt{V(m)}}$$

ถ้าตัวสถิติ Z มีนัยสำคัญก็แสดงว่าอนุกรมข้อมูลนั้นไม่เป็นอิสระกัน แต่ถ้าไม่มีนัยสำคัญก็แสดงว่าเป็นอิสระกัน

ตัวอย่าง จากตัวอย่างที่แล้ว ๆ มา จะได้

$R = 9 = R(+) \quad R_2 = R(-) = 10 \quad R_3 = R(0) = 0 \quad R = 9+10+0 = 19$

$m = \{26(27) - (9^2+10^2+0^2)\}/26 = 20.038$

$$V(m) = \frac{181\{181+26(27)\} - 2(26)(1729) - 26^3}{26^3(26-1)} = 3.0969822$$

$\Sigma n_i^2 = 9^2+10^2+0^2 = 181$ และ $\Sigma n_i^3 = 9^3+10^3+0^3 = 1729$

$$Z = \frac{(19-0.5) - 20.038}{\sqrt{3.097}} = -0.31$$

สำหรับ $\alpha = .05$ ค่าวิกฤต $Z_{.025} = 1.96$ จึงสรุปได้ว่า อนุกรมข้อมูลเป็นอิสระกัน

7.2 แบบทดสอบสหสัมพันธ์อันดับ (Rank - Correlation Test)

สหสัมพันธ์อันดับของสเปียร์แมน หรือเคนดาลล์ นั้นใช้ทดสอบการสุ่มได้โดยทดสอบนัยสำคัญระหว่างคู่ (i, R_i); i = 1, 2, .., n ในเมื่อ R_i เป็นอันดับที่ของค่าสังเกต X_i ที่เรียงตามขนาด ถ้าสหสัมพันธ์อันดับมีนัยสำคัญก็แสดงว่า ไม่เป็นแบบสุ่ม

ตัวอย่าง ในการทดสอบว่าอนุกรมต่อไปนี้เป็นแบบสุ่มหรือไม่โดยอาศัยแบบทดสอบสหสัมพันธ์อันดับ ทำได้ดังนี้

i	X_i	R_i	i	X_i	R_i	i	X_i	R_i
1	-0.49	17	8	-0.88	19	15	0.75	9
2	-0.92	20	9	1.56	4	16	2.84	1
3	-0.22	16	10	-0.11	15	17	0.94	8
4	0.01	14	11	0.29	12	18	1.64	3

i	X_i	R_i	i	X_i	R_i	i	X_i	R_i
5	0.47	11	12	1.18	5	19	2.32	2
6	0.67	10	13	1.15	6	20	1.03	7
7	-0.78	18	14	0.10	13			

$$r_s = 1 - 6\Sigma d_i^2/(n^3-n)$$

$$= 1 - 6(2318)/(20^3-20) = -0.7429$$

ในเมื่อ $d_i = i - R_i$; $i = 1, 2, .., 20$

ตัวสถิติทดสอบที่ใช้ทดสอบนัยสำคัญของ r_s ก็คือ

$$T = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}}$$

โดยมีค่าเป็น $T = -0.7429/\sqrt{(1-(-0.7429)^2)/(20-2)}$

$= -4.208$

แสดงว่า T มีนัยสำคัญอย่างยิ่ง นั่นคืออนุกรมข้อมูลนั้นไม่เป็นแบบสุ่ม

7.3 แบบทดสอบอัตราส่วนของ ฟอน นิวแมนน์ (Von Neumann Ratio Test)

แบบทดสอบนี้ใช้ทดสอบอนุกรมของค่าสังเกตว่าเป็นแบบสุ่มหรือไม่โดยอาศัยผลต่างที่ติดกัน ดังนั้นแบบทดสอบนี้จึงได้ชื่ออีกอย่างว่า แบบทดสอบกำลังสองเฉลี่ยของผลต่างที่ติดกัน

ให้ $X_1, X_2, .., X_n$ เป็นอนุกรมค่าสังเกตซึ่งเรียงตามเกณฑ์อย่างหนึ่ง (ตามปฏิทิน) ถ้า X_i ขึ้นอยู่กับ X_{i-1} $(i = 1, 2, .., n)$ นั่นคืออนุกรม X_i ไม่เป็นเชิงสุ่ม แล้วอนุกรมค่าสังเกตจะให้รูปแบบ (Pattern) บางอย่าง ในประชากร เราเรียกความไม่เป็นเชิงสุ่มหรือความพึ่งพิงของอนุกรมค่าสังเกตว่า สหสัมพันธ์ในตัวเอง (Autocorrelation) แต่ในตัวอย่างเรียกว่า สหสัมพันธ์เชิงอนุกรม (Serial Correlation) แต่สองคำนี้มักจะใช้แทนกัน เมื่อ X_i ขึ้นอยู่กับ X_{i-1} แล้วผลต่างที่ติดกัน หรือ (X_i-X_{i-1}) จะมีค่าน้อย เพื่อที่จะขจัดเครื่องหมายของผลต่างนี้ จึงใช้กำลังสองเฉลี่ยของผลต่างที่ติดกัน นั่นคือ

$$\delta^2 = \Sigma(X_i - X_{i-1})^2/(n-1)$$

ถ้าอนุกรมของ X_i เป็นแบบสุ่ม (ซึ่งเป็นสมมติฐานหลัก Ho) แล้วจะได้ว่า

$$E(\delta^2) = 2\sigma^2$$

แต่ไม่ทราบความแปรปรวนประชากร σ^2 จึงประมาณด้วยความแปรปรวนตัวอย่าง

$$S_o^2 = \Sigma(X_i - \overline{X})^2/n$$

ซึ่งเป็นค่าประมาณที่เอียงเฉ

ดังนั้นถ้ากำหนดอัตราส่วนฟอน นิวแมนน์ (VNR) ไว้เป็น

$$VNR = \delta^2/S_o^2$$

แล้วค่าคาดหวังของ VNR จะประมาณ 2

ถ้า X แจกแจงแบบปกติ และถ้า n > 60 แล้ว VNR จะแจกแจงใกล้กับการแจกแจงปกติ ด้วยค่าเฉลี่ย และความแปรปรวนเป็น

$$E(VNR) = 2n/(n-1)$$

$$V(VNR) = 4n^2(n-2)/(n+1)(n-1)^3$$

เมื่อ n โตมาก $E(VNR) \cong 2$ และ $V(VNR) \cong 4/n$

สำหรับตัวอย่างขนาดเล็ก ($n \leq 60$) ค่าวิกฤตพิจารณาจากตารางพิเศษ แต่ถ้าตัวอย่างขนาดโต ก็ใช้ตัวสถิติ Z ซึ่งมีการแจกแจงปกติมาตรฐาน

$$Z = \frac{VNR - E(VNR)}{V(VNR)}$$

ตัวอย่าง ในการศึกษารายได้จากการขายสินค้าชนิดหนึ่งใน 50 วัน ติดต่อกัน ปรากฏว่ารายได้สรุปได้ดังนี้

$$\Sigma X = 4864 \qquad \Sigma X^2 = 598628$$

$$\Sigma(X_i - X_{i-1})^2 = 273993$$

รายได้เป็นแบบสุ่มหรือไม่ ?

Ho : รายได้ต่อวันเป็นแบบสุ่ม

Ha : รายได้ต่อวันไม่เป็นแบบสุ่ม

$$\delta^2 = \Sigma(X_i - X_{i-1})^2/(n-1) = 273993/49 = 5591.69$$

$$S_o^2 = \Sigma(X_i - \bar{X})^2/n = (\Sigma X^2 - (\Sigma X)^2/n)/n$$

$$= (598628 - (4864)^2/50)/50 = 2509.16$$

$$VNR = \delta^2/S_o^2 = 5591.69/2509.16 = 2.229$$

จากตารางพิเศษ จะสรุปได้ว่า รายได้เป็นแบบสุ่ม

เมื่อใช้ตัวสถิติ Z จะได้ค่าเป็น

$$Z = \frac{2.229 - 2}{\sqrt{4/50}} = 0.8096$$

ซึ่งจะสรุปได้เช่นเดียวกัน นั่นคือรายได้ต่อวันเป็นแบบสุ่ม

นอกจากแบบทดสอบดังกล่าวนี้ ยังมีแบบทดสอบแนวโน้มของคอกซ์และสจวร์ท ที่ใช้สำหรับทดสอบการสุ่มของอนุกรมได้ ซึ่งจะได้กล่าวต่อไปในปัญหาอื่นๆ

THE LEAST INITIAL DEVIATION FROM THE TRUTH IS MULTIPLIED LATER A THOUSANDFOLD.

ARISTOTLE