บทที่ 2

ตัวสถิติและผลงานวิจัยที่เกี่ยวข้อง

ตัวสถิติที่ใช้ในการทดสอบการแจกแจงปกติของประชากรมีหลายแบบ ซึ่งได้มีการพัฒนาขึ้นเรื่อย ๆ พร้อมทั้งได้มีผลงานวิจัยศึกษาเปรียบเทียบอำนาจการทดสอบของตัวสถิติต่าง ๆ ในบทนี้จะกล่าวถึงรายละเอียดของตัวสถิติที่ศึกษาครั้งนี้ พร้อมทั้งยกตัวอย่างการใช้ตัวสถิติ และตอนท้ายของบทจะนำเสนอผลงานวิจัยที่เกี่ยวข้อง ซึ่งรายละเอียดต่าง ๆ เป็นดังนี้

2.1 ตัวสถิติที่ใช้ในการทดสอบ

2.1.1 ตัวสถิติ χ^2 หรือ การทดสอบไคสแควร์ (Chi-Square Tests Goodness of Fit of Normality)

Karl Pearson เป็นผู้เริ่มใช้การแจกแจงไคสแควร์ ในการทดสอบการสอดคล้องระหว่างค่าสังเกตและสมมติฐาน โดยใช้ข้อมูลที่ได้จากการนับหรืออยู่ในรูปของความถี่ ในปัจจุบันตัวสถิติ χ^2 เป็นตัวสถิติทดสอบที่นิยมใช้ในการทดสอบการแจกแจงปกติมาก ถึงแม้ว่าการทดสอบไคสแควร์จะมีปัญหาในเรื่องเกี่ยวกับขนาดตัวอย่างและการแบ่งชั้นให้เหมาะสมแก่ตัวอย่าง

วิธีการในการทดสอบไคสแควร์มีดังนี้คือ ขั้นแรกแบ่งกลุ่มข้อมูลตัวอย่างเป็นกลุ่มที่แยกจากกันโดยเด็ดขาด หรือแบ่งตามช่วงค่า และนับจำนวนข้อมูลหรือความถี่อยู่ในแต่ละช่วง (Observed Frequencies) ต่อจากนั้นก็ใช้หลักการทางการแจกแจงแบบปกติ คำนวณค่าความถี่ที่คาดหวัง (Expected Frequencies) ในแต่ละช่วงค่า ถ้ามีความแตกต่างมากระหว่างจำนวนที่สังเกตกับค่าที่คาดหวัง สรุปว่าข้อมูลไม่ได้มาจากการแจกแจงปกติ ถ้าแตกต่างกันไม่มากก็สรุปว่าตัวอย่างมาจากการแจกแจงปกติ สูตรในการคำนวณ

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

โดยที่ X^2 มีการแจกแจงแบบไคส์แควร์ที่องศาอิสระเท่ากับ k-m-1 เมื่อ m คือ จำนวนพารามิเตอร์ใน Model ตามสมมติฐานที่ถูกประมาณจากข้อมูล ในที่นี้เท่ากับ 2 และองศาอิสระที่หายไปอีก 1 เนื่องจากข้อจำกัดที่ว่า $\sum_{i=1}^{k} O_i = n = \sum_{i=1}^{k} E_i$

n = ขนาดของตัวอย่าง

k = จำนวนช่วงของข้อมูล

O_i = จำนวนความถี่ของข้อมูลในช่วงที่ i คือ x_{i-1} ถึง x_i และเป็นตัวแปรที่ไม่ขึ้นต่อกัน (Mutually Independent Random Variable)

E_i = ค่าคาดหวังของจำนวนความถี่ของข้อมูลในช่วงที่ i โดยค่าคาดหวังนี้คำนวณได้จากความน่าจะเป็นที่คาดหวังไว้ คูณกับขนาดตัวอย่าง

ซึ่ง $E_i = \text{Prob}\left(\frac{x_{i-1} - \bar{x}}{s} \leq z \leq \frac{x_i - \bar{x}}{s}\right) \times n$

$$\bar{x} = \frac{\sum_{i=1}^{k} O_i \bar{x}_i}{\sum_{i=1}^{k} O_i}$$ (เมื่อ $\bar{x}_i$ จุดกึ่งกลางของช่วงข้อมูลที่ i)

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{k} O_i \bar{x}_i^2 - (n\bar{x}^2)\right)$$

จะปฏิเสธสมมติฐานว่าประชากรมีการแจกแจงปกติ เมื่อค่าสถิติที่คำนวณได้มากกว่า χ^2 ที่องศาอิสระเท่ากับ k-3

ข้อจำกัดของตัวสถิติ

1) $E_i \geq 5$

2) ถ้า $E_i < 5$ อนุโลมได้เมื่อ $1 < E_i < 5$ ไม่เกิน 20% ของจำนวนช่วงของข้อมูล วิธีแก้ไขสำหรับ $E_i < 5$ เกิน 20% และ $E_i < 1$ คือการรวมข้อมูลข้อมูลขั้นที่อยู่ติดกันเข้าไปด้วย เพื่อทำให้สามารถใช้สูตรได้

<u>หมายเหตุ</u> การทดสอบไคสแควร์เป็นการทดสอบแบบหางเดียว ทางด้านมากกว่าเสมอ ทั้งนี้เพราะว่า O_i จะต่างกัน E_i อย่างมีนัยสำคัญ ถ้า $(O_i - E_i)^2$ มีค่ามาก ซึ่งได้มาจะมีเฉพาะค่าบวกอย่างเดียว

<u>ตัวอย่างที่ 1</u> การใช้การทดสอบไคสแควร์ในการทดสอบการแจกแจงปกติ

ข้อมูลข้างล่างนี้เป็นการแจกแจงความสูงของนักศึกษา 100 คน ต้องการทดสอบโดยใช้ระดับความมีนัยสำคัญ .05 ดูว่า การแจกแจงนี้เป็นแบบปกติหรือไม่

ความสูง (x เป็นนิ้ว)	60-62	63-65	66-68	69-71	72-74
จำนวนนักศึกษา	5	18	42	27	8

H_o : การแจกแจงความสูงของนักศึกษาเป็นแบบปกติ

H_1 : การแจกแจงความสูงของนักศึกษาไม่เป็นแบบปกติ

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= \sum_{i=1}^{k} \frac{(O_i^2 - 2O_iE_i + E_i^2)}{E_i}$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{E_i} - 2 \sum_{i=1}^{k} \frac{O_i E_i}{E_i} + \sum_{i=1}^{k} \frac{E_i^2}{E_i}$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n \quad (\because \sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n)$$

ต้องหาค่า E_i ก่อน ในการหาค่า E นั้นต้องใช้พื้นที่ใต้โค้งปกติ จึงต้องปรับค่าให้เป็นแบบต่อเนื่องกัน แล้วเปลี่ยนจากค่า x เป็น z เพื่อเปิดหาพื้นที่จากตารางปกติได้

อันตรภาคชั้น	O_i	ขีดจำกัดบน(x_i)	$x_i-\bar{x}$	$z = \frac{x_i-\bar{x}}{S}$	ความน่าจะเป็น (จากตาราง Z)	E_i	$\frac{O_i^2}{E_i}$
60-62	5	62.5	-4.95	-1.70	.0446	4.46	5.6
63-65	18	65.5	-1.95	-0.67	.2068	20.68	15.6
66-68	42	68.5	1.05	0.36	.3892	38.92	45.3
69-71	37	71.5	4.05	1.39	.2771	27.71	26.3
72-74	8	74.5	7.05	2.41	.0823	8.23	7.8
	100				1.0000	100	100.6

เนื่องจาก $z = \frac{x - \mu}{\sigma}$ ในที่นี้ μ และ σ ไม่ทราบต้องประมาณมาจากข้อมูลในตัวอย่าง ซึ่งจะได้ $\bar{x}$ = 67.45 นิ้ว และ s = 2.92 นิ้ว

การคำนวณความถี่คาดหวังทำดังนี้คือ

ในอันตภาคชั้น 63-65 ซึ่งมีความถี่ที่สังเกตได้ 18 ขีดจำกัดของอันตรภาคชั้นนี้เป็น 62.5-65.5

$$\therefore P(62.5 \leqslant x \leqslant 65.5) = P\left(\frac{62.5 - 67.5}{2.92} \leqslant z \leqslant \frac{65.5 - 67.5}{2.92}\right)$$

$$= P(-1.70 \leqslant z \leqslant -0.67)$$

$$= P(z \leqslant -0.67) - P(z \leqslant -1.70)$$

$$= 0.2514 - 0.0446$$

$$= 0.2068$$

$\therefore$ ความถี่คาดหวัง (E_i) $= 0.2068 \times 100 = 20.68$

$$\therefore \quad \chi^2 = 100.6 - 100 = 0.6$$

แต่ค่าของ χ^2 จากตารางแจกแจงความน่าจะเป็นแบบ χ^2 ที่ $[(5-1)-2 = 2]$ องศาอิสระ และระดับนัยสำคัญ 0.05 เท่ากับ 5.99

ดังนั้นยอมรับ H_o ที่ว่าความสูงของนักศึกษามีการแจกแจงแบบปกติ

2.1.2 ตัวสถิติ u หรือ Studentized Range Test

$$u = \frac{x_{(n)} - x_{(1)}}{s} \quad \text{หรือ} \quad \frac{w}{s}$$

เมื่อ $x_{(i)}$ = order sample

$$s^2 = \sum_{i=1}^{n} \frac{(x_{(i)} - \bar{x})^2}{n-1}$$

$$\bar{x} = \sum_{i=1}^{n} \frac{x_{(i)}}{n}$$

n = ขนาดตัวอย่าง

ค่าวิกฤตของตัวสถิติ u ใช้ตารางของ Pearson และ Stephens (1964:486) (ดูจากตารางที่ 1 ในภาคผนวก) จะปฏิเสธสมมติฐานว่าประชากรไม่มีการแจกแจงปกติ เมื่อค่า u ที่คำนวณได้มีค่ามากกว่า Upper Percentage Point หรือน้อยกว่า Lower Percentage Point จากตาราง

ตัวสถิติ u เป็นตัวสถิติที่คิดขึ้นโดย David, Hartly และ Pearson (1954) เพื่อใช้ในการทดสอบ Homogeneity และการแจกแจงปกติของข้อมูล โดยทำการศึกษาต่อจาก G.A. Baker (1946:366) ซึ่งประสบผลสำเร็จในการใช้สัดส่วนของ $\frac{w}{s}$ ในการทดสอบ Homogeneity ของข้อมูลและเลียนแบบการศึกษาของ Geary (1935) ซึ่งศึกษาสัดส่วนของส่วนเบี่ยงเบนของค่าเฉลี่ยหารด้วยส่วนเบี่ยงเบนมาตรฐาน เทียบกับ b_2 (Standard Fourth Moment) เพื่อใช้ในการทดสอบการแจกแจงปกติ โดย David และคณะได้ทำการศึกษาเปรียบเทียบระหว่าง b_2 และ $\frac{w}{s}$ ในลักษณะตัวอย่างและการแจกแจงหลาย ๆ ลักษณะ ผลการศึกษาพบว่า b_2 และ $\frac{w}{s}$ มีความสัมพันธ์สูง โดยเฉพาะในการแจกแจงที่มีลักษณะ Platykurtic และเมื่อเปรียบเทียบอำนาจในการทดสอบการแจกแจงปกติ พบว่า $\frac{w}{s}$ ใช้ได้ผลดีพอกับ b_2 และดีกว่าในบางกรณีด้วย

<u>ตัวอย่างที่ 2</u> การใช้ Studentized Range test ในการทดสอบการแจกแจงปกติ

ข้อมูลตัวอย่าง	y_1	y_2	y_3	y_4	y_5	y_6	y_7
	6	1	-4	8	-2	5	0

H_o : ประชากรมีการแจกแจงแบบปกติ

H_1 : ประชากรไม่มีการแจกแจงแบบปกติ

เรียงลำดับข้อมูลใหม่

i	$x_{(i)}$	$(x_{(i)} - \bar{x})^2$
1	-4	36
2	-2	16
3	0	4
4	1	1
5	5	9
6	6	16
7	8	36
	14	118

$$\therefore \quad \bar{x} = \sum_{i=1}^{7} \frac{x_{(i)}}{7} = \frac{14}{7} = 2$$

$$s^2 = \sum_{i=1}^{7} \frac{(x_{(i)} - \bar{x})^2}{7-1}$$

$$= \frac{118}{6} = 19.667$$

$$\therefore \quad u = \frac{x_{(n)} - x_{(1)}}{s}$$

$$= \frac{8 - (-4)}{4.435} = 2.706$$

จากตารางที่ 1 ในภาคผนวกพบว่าค่า u ที่ $n = 7$ ที่ระดับนัยสำคัญ 0.01 คือ $[2.32 , 3.369]$ ดังนั้นปฏิเสธ H_o นั่นคือประชากรมีการแจกแจงแบบปกติ

2.1.3 ตัวสถิติ W หรือตัวสถิติของ Shapiro และ Wilk (Shapiro and Wilk Statistic)

$$W = \frac{\left\{ \sum_{i=1}^{k} a_{n-i+1} \left(x_{(n-i+1)} - x_{(i)}\right) \right\}^2}{\sum_{i=1}^{n} \left(x_{(i)} - \bar{x}\right)^2}$$

เมื่อ n = ขนาดตัวอย่าง

k = จำนวนเต็มที่เล็กที่สุดที่มากกว่าหรือเท่ากับ $\frac{n}{2}$

a_i = ค่าสัมประสิทธิ์ซึ่งเปิดจากตารางที่ 2 ในภาคผนวก เมื่อ $n \leqslant 50$

$x_{(i)}$ = order sample

จะปฏิเสธสมมติฐานว่าง (H_o : ประชากรมีการแจกแจงปกติ) เมื่อค่าสถิติที่คำนวณได้น้อยกว่า W ที่เปิดจากตารางที่ขนาดตัวอย่าง n และระดับนัยสำคัญที่ต้องการ

ตัวสถิติ W เป็นตัวสถิติที่ใช้สัดส่วนของ $\hat{\sigma}^2/s^2$ เมื่อ $\hat{\sigma}^2$ เป็นค่าประมาณของ σ^2 โดยมีพื้นฐานจาก

ถ้ากำหนดให้ $m' = (m_1 , m_2, \ldots, m_n)$ เป็นเวคเตอร์ของค่าคาดหวังของ $N(0,1)$

$v = (v_{ij})$ เป็น Covariance Matrix ขนาด $n \times n$

ถ้า $y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n$ เป็น Order sample ขนาด n จาก $N(0, 1)$

ดังนั้น $E(y_i) = m_i$

$Cov(y_i , y_j) = v_{ij}$ $(i,j = 1,2,\ldots, n)$

ให้ $x' = (x_1, x_2, \ldots, x_n)$ เป็นเวคเตอร์ของ order sample ของตัวอย่างขนาด n ต้องการทดสอบดูว่าตัวอย่างนี้สุ่มมาจากการแจกแจงปกติ $N(\mu, \sigma^2)$

ถ้า $\{x_i\}$ มีการแจกแจงปกติ ดังนั้น

$$\frac{x_i - \mu}{\sigma} = y_i$$

$$x_i = \mu + \sigma y_i \qquad (i = 1,2,\ldots, n)$$

$$E(x_i) = \mu + \sigma E(y_i)$$

$$E(x) = \mu 1 + \sigma m \qquad (1 = 1_{nx1})$$

ดังนั้น $$E(x) = p\theta \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

เมื่อ P เป็นเมตริกซ์ของ $(1, m_i)$ ซึ่งมีขนาด nx2

θ' เป็นเวคเตอร์ของ (μ, σ)

และ $\mathrm{Var}(x_i) = \sigma^2 \mathrm{Var}(y_i)$

$\mathrm{Var}(x) = \sigma^2 v$

จาก (1) โดยทฤษฎี Generalized Least-Squares

$$\hat{\theta} = (p' v^{-1} p)^{-1} p' v^{-1} x \qquad \ldots\ldots\ldots\ldots \quad (2)$$

พิจารณา $p' v^{-1} p$ สามารถกระจายในรูปของ

$$p' v^{-1} p = \begin{bmatrix} 1' v^{-1} 1 & 1' v^{-1} m \\ 1' v^{-1} m & m' v^{-1} m \end{bmatrix}$$

$$\therefore \quad (p' v^{-1} p)^{-1} = \frac{1}{\Delta} \begin{bmatrix} m' v^{-1} m & -1' v^{-1} m \\ -1' v^{-1} m & 1' v^{-1} 1 \end{bmatrix}$$

เมื่อ $\Delta = 1'V^{-1}1m'V^{-1}m - 1'V^{-1}m\,1'V^{-1}m$

$$= 1'V^{-1}1\,m'V^{-1}m - (1'V^{-1}m)^2$$

ในทำนองเดียวกันสามารถกระจาย $p'V^{-1}x$ ในรูปของ

$$p'V^{-1}x = \begin{bmatrix} 1'V^{-1}x \\ m'V^{-1}x \end{bmatrix}$$

จาก (2) ดังนั้น

$$\hat{\theta} = \frac{1}{\Delta}\begin{bmatrix} m'V^{-1}m & -1'V^{-1}m \\ -1'V^{-1}m & 1'V^{-1}1 \end{bmatrix}\begin{bmatrix} 1'V^{-1}x \\ m'V^{-1}x \end{bmatrix}$$

$$\therefore \begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix} m'V^{-1}m\,1'V^{-1}x - 1'V^{-1}mm'V^{-1}x \\ -1'V^{-1}m1'V^{-1}x + 1'V^{-1}1m'V^{-1}x \end{bmatrix}$$

$$\therefore \quad \hat{\mu} = \frac{m'V^{-1}(m1' - 1m')V^{-1}x}{1'V^{-1}1m'V^{-1}m - (1'V^{-1}m)^2}$$

$$\therefore \quad \hat{\sigma} = \frac{1'V^{-1}(1m' - m1')V^{-1}x}{1'V^{-1}1\,m'V^{-1}m - (1'V^{-1}m)^2} \qquad (\because 1'V^{-1}m = 0)$$

$$= \frac{m'V^{-1}x}{m'V^{-1}m}$$

เนื่องจาก $s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2$ เป็นตัวประมาณไม่เอนเอียงของ $(n-1)\,\sigma^2$
ดังนั้นตัวสถิติ W ในการทดสอบการแจกแจงปกติ คือ

014224

$$W = \frac{R^4 \hat{\sigma}^2}{C^2 S^2} = \frac{b^2}{S^2} = \frac{(a' X)^2}{S^2} = \frac{(\Sigma a_i x_i)^2}{\Sigma (x_i - \bar{x})^2}$$

เมื่อ $R^2 = m' V^{-1} m$

$C^2 = m' V^{-1} V^{-1} m$

$a' = (a_1, a_2, \ldots, a_m) = \dfrac{m' V^{-1}}{(m' V^{-1} V^{-1} m)^{\frac{1}{2}}} = \dfrac{m' V^{-1}}{C}$

$b = \dfrac{R^2 \hat{\sigma}}{C}$

แต่เนื่องจาก $- a_i = a_{n-i+1}$ จาก Shapiro และ Wilk (1965 : 593)

ดังนั้น

$$W = \frac{\left\{ \sum_{i=1}^{k} a_{n-i+1} \left(x_{(n-i+1)} - x_{(i)} \right) \right\}^2}{\sum_{i=1}^{n} \left(x_{(i)} - \bar{x} \right)^2}$$

Sarhan และ Greenberg ได้คำนวณค่าของ a' เมื่อขนาดตัวอย่างเล็กกว่า 20 สำหรับขนาดตัวอย่างที่ใหญ่กว่า 20 Shapiro และ Wilk ได้ประมาณค่า a' โดย

$$a' = \frac{m' V^{-1}}{(m' V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

$\therefore \quad a' a = 1$

ให้ $a^* = m' V^{-1}$

ดังนั้น $c^2 = a^* a^{*\prime}$

และประมาณค่า a^* โดย $\hat{a}_i^* = 2 m_i$ เมื่อ $i = 2, 3, \ldots, n-1$

และ $$(\hat{a}_1^*)^2 = (\hat{a}_n^*)^2 = \begin{cases} \dfrac{\tau(\frac{1}{2}n)}{\sqrt{2\,\tau\{\frac{1}{2}(n+1)\}}} & (n < 20) \\ \dfrac{\tau\{\frac{1}{2}(n+1)\}}{\sqrt{2\quad\tau(\frac{1}{2}n+1)}} & (n > 20) \end{cases}$$

<u>ตัวอย่างที่ 3</u> การใช้ตัวสถิติของ Shapiro และ Wilk ในการทดสอบการแจกแจงปกติ

ข้อมูลตัวอย่าง

y_1	y_2	y_3	y_4	y_5	y_6	y_7
6	1	-4	8	-2	5	0

H_o : ประชากรมีการแจกแจงแบบปกติ

H_1 : ประชากรไม่มีการแจกแจงแบบปกติ

เรียงลำดับข้อมูลใหม่

i	$x_{(i)}$	$(x_{(i)} - \bar{x})^2$
1	-4	36
2	-2	16
3	0	4
4	1	1
5	5	9
6	6	16
7	8	36
	14	118

$$\therefore \quad S^2 = \sum_{i=1}^{7} (x_{(i)} - \bar{x})^2$$

$$= \quad 118$$

จากตารางที่ 2 ในภาคผนวก เมื่อ $n = 7$

$$a_7 = 0.6233 \; , \; a_6 = 0.3031 \; , \; a_5 = 0.1401 \; , \; a_4 = 0.0000$$

$$\therefore \quad b = \sum_{i=1}^{k} a_{n-i+1} (x_{(n-i+1)} - x_{(i)})$$

$$= \quad 0.6233(8+4) + 0.3031(6+2) + 0.1401(5-0)$$

$$= \quad 10.6049$$

$$W = \frac{b^2}{S} = \frac{(10.6049)^2}{118} = 0.9530$$

จากตารางที่ 3 พบว่าค่า W ที่ $n = 7$ และระดับนัยสำคัญ 0.01 คือ 0.730 ดังนั้น ยอมรับ H_0 เนื่องจากค่า W ที่คำนวณมีค่ามากกว่า 0.730

2.1.4 ตัวสถิติ r (Probability Plot Correlation Coefficient Test) หรือตัวสถิติของ Felliben (Felliben Statistic)

$$r = \mathrm{Cor}\,(x\ ,\ m)$$

$$= \frac{\sum_{i=1}^{n} (x_{(i)} - \bar{x})(m_{(i)} - \bar{m})}{\sqrt{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 \sum_{i=1}^{n} (m_{(i)} - \bar{m})^2}}$$

เมื่อ $x_{(i)}$ = order sample

$m_{(i)}$ = order statistic medians $m_{(i)}$ จาก $N(0,1)$

$= \phi^{-1}(M_i)$

ซึ่ง

$$M_i = \begin{cases} 1 - M_n & i = 1 \\ \dfrac{(i-0.3175)}{(n+0.365)} & i = 2,3,\ldots,\ n-1 \\ 0.5^{(1/n)} & i = n \end{cases}$$

n = ขนาดตัวอย่าง

จะปฏิเสธสมมติฐานว่าง (H_o : ประชากรมีการแจกแจงปกติ) เมื่อค่าที่คำนวณได้น้อยกว่าค่า r ที่เปิดจากตารางที่ 4 ในภาคผนวกตามขนาดตัวอย่าง n และระดับนัยสำคัญที่ศึกษา

ตัวสถิติ r เป็นตัวสถิติที่ James J. Filliben เป็นผู้คิดขึ้นในปี ค.ศ. 1975 โดยใช้สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ระหว่าง order sample กับ order statistic median จากการแจกแจงปกติ N(0,1) Felliben ได้ดัดแปลงตัวสถิติของ Shapiro และ Wilk ซึ่งศึกษาโดยใช้ Probability Plot* ระหว่าง $x_{(i)}$ ซึ่งเป็น order sample กับ $E(x_i)$ (จากหัวข้อ 2.1.3 $W = \frac{(a'x)^2}{S^2}$ และ $a' = \frac{m'v^{-1}}{C^2}$ ซึ่งคำนวณได้จาก $E(x_i)$) จากการศึกษาของ Filliben พบว่าการหาค่าของ $E(x_i)$ มีความยุ่งยากในแง่ของเทคนิคอินติเกรต และในบางครั้ง $E(x_i)$ ก็ไม่สามารถคำนวณได้

Filliben จึงดัดแปลงตัวสถิติให้ง่ายขึ้น โดยใช้ Probability Plot ระหว่าง $x_{(i)}$ กับ loc $(x_{(i)})$ = med $(x_{(i)})$ = $m_{(i)}$ ซึ่ง $m_{(i)}$ จะเป็น order statistic median จากการแจกแจงปกติ N(0,1) ถ้าสมมติฐานว่างถูกต้อง จุดระหว่าง $x_{(i)}$ กับ $m_{(i)}$ จะเข้าใกล้เส้นตรง ดังนั้น Filliben ซึ่งใช้สัมประสิทธิ์สหสัมพันธ์ ในการจัดว่าจุดระหว่าง $x_{(i)}$ กับ $m_{(i)}$ จะเข้าใกล้เส้นตรงหรือไม่ ตัวสถิติที่ได้จึงเป็น

$$r = \text{Cor}(x, m)$$

$$= \frac{\sum_{i=1}^{n} (x_{(i)} - \bar{x})(m_{(i)} - \bar{m})}{\sqrt{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 \sum_{i=1}^{n} (m_{(i)} - \bar{m})^2}}$$

---

*Probability Plot หมายถึง การ plot จุดระหว่างค่าสังเกตกับค่าคาดหวังภายใต้สมมติฐาน

เมื่อ $m_{(i)} = \phi^{-1}(M_i)$

$$M_i = \begin{cases} 1 - M_n \\ \dfrac{(1 - 0.3175)}{(n + 0.365} \\ 0.5^{(1/n)} \end{cases}$$

แต่เนื่องจาก $m_{(i)} = -m_{(n-i+1)}$

$$\therefore \quad \bar{m} = 0$$

$$\therefore \quad r = \frac{\sum_{i=1}^{n} (x_{(i)} - \bar{x})\, m_{(i)}}{\sqrt{\sum_{i=1}^{n} (x_{(i)} - \bar{x}) \sum_{i=1}^{n} (m_{(i)} - \bar{m})^2}}$$

$$= \frac{\sum_{i=1}^{n} x_{(i)} m_{(i)} - \bar{x} \sum_{i=1}^{n} m_{(i)}}{\sqrt{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 \sum_{i=1}^{n} (m_{(i)} - \bar{m})^2}}$$

$$= \frac{\sum_{i=1}^{n} x_{(i)} m_{(i)}}{\sqrt{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 \sum_{i=1}^{n} (m_{(i)} - \bar{m})^2}} \qquad \left(\because \sum_{i=1}^{n} m_{(i)} = 0\right)$$

<u>ตัวอย่างที่ 4</u> การใช้ตัวสถิติของ Filliden ในการทดสอบการแจกแจงปกติ

ข้อมูลตัวอย่าง

y_1	y_2	y_3	y_4	y_5	y_6	y_7
6	1	4	8	-2	5	0

H_o : ประชากรมีการแจกแจงแบบปกติ

H_1 : ประชากรไม่มีการแจกแจงแบบปกติ

ลำดับ(i)	$y_{(i)}$	$x_{(i)}$	M_i	$m_{(i)} = \phi^{-1}(M_i)$	$m^2_{(i)}$	$(x_{(i)}-\bar{x})^2$	$m_{(i)}x_{(i)}$
1	6	-4	.0943	-1.31493	1.72904	36	5.25972
2	1	-2	.2284	-0.74388	0.55334	16	1.48776
3	4	0	.3642	-0.34681	0.12028	4	0
4	8	1	.5000	0	0	1	0
5	-2	5	.6358	0.34681	0.12028	9	1.73405
6	5	6	.7716	0.74388	0.55336	16	4.46328
7	0	8	.9057	1.31493	1.72904	36	10.51944
					4.80535	118	23.46425

$$m_{(1)} = \phi^{-1}(.0943)$$

จากตารางการแจกแจงปกติมาตรฐานพบว่า $m_{(1)}$ อยู่ระหว่าง $z = -1.31$ และ $z = -1.32$

$$\therefore \; m_{(1)} = -1.31493$$

$$r = Cor(x,m)$$

$$= \frac{\sum_{i=1}^{7} m_{(i)} x_{(i)}}{\sqrt{\sum_{i=1}^{7} m^2_{(i)} \sum_{i=1}^{7} (x_{(i)} - \bar{x})^2}}$$

$$= \frac{23.46425}{\sqrt{(4.80535)(118)}}$$

$$= 0.98538$$

จากตารางที่ 4 เมื่อ $n = 7$ พบว่า r อยู่ระหว่าง 75% และ 90% (p value มากกว่าα) ดังนั้นยอมรับ H_o เป็นจริงนั้นคือ ประชากรมีการแจกแจงปกติ

2.1.5 ตัวสถิติ T_1' และ T_2' หรือ ตัวสถิติของ Hannu Oja (Hannu Oja Statistics)

Hannu Oja ได้สนใจศึกษาการทดสอบการแจกแจงปกติ โดยเสนอสถิติทดสอบถึง 2 ครั้งคือในปี ค.ศ. 1981 Oja เสนอตัวสถิติที่ดัดแปลงมาจากตัวสถิติของ David และ Quade(1978) เพื่อใช้ในการทดสอบสมมติฐานหลายพันธ์ (Composite Hypothesis) ตัวสถิติที่เสนอคือ

$$T_1 = \binom{n}{3}^{-1} \sum \frac{x_{(k)} - x_{(j)}}{x_{(k)} - x_{(i)}} \quad , \quad 1 \leqslant i < j < k \leqslant n$$

$$T_2 = \binom{n}{4}^{-1} \sum \frac{x_{(k)} - x_{(j)}}{x_{(\ell)} - x_{(i)}} \quad 1 \leqslant i < j < \ell \leqslant n$$

เมื่อ $x_{(i)}$ คือ order sample และใช้ในการทดสอบความเบ้และความโด่งตามลำดับค่าวิกฤกติของตัวสถิติประมาณด้วยการแจกแจงปกติ

ค.ศ. 1983 Oja ได้เสนอตัวสถิติตัวใหม่ในการทดสอบการแจกแจงปกติ (New Tests Normality) โดยทำการพัฒนาจากตัวสถิติที่เสนอไว้ ตัวสถิติตัวใหม่ที่เสนอคือ

$$T_1' = \binom{n}{3}^{-1} \Sigma \log \frac{x_{(k)} - x_{(j)}}{x_{(j)} - x_{(i)}} \quad , \quad 1 \leqslant i < j < k \leqslant n$$

$$= \Sigma\Sigma \; a_{ij} \log \left(x_{(j)} - x_{(i)}\right), \; 1 \leqslant i < j \leqslant n$$

เมื่อ $x_{(i)}$ = order sample

$$a_{ij} = \binom{n}{3}^{-1} (i + j - n-1)$$

T_1' มีความไวต่อความเบ้

$$T_2' = \binom{n}{4}^{-1} \Sigma \log \left[\frac{(x_{(k)} - x_{(j)})^2}{(x_{(\ell)} - x_{(k)})(x_{(j)} - x_{(i)})} \right] , \; 1 \leqslant i < j < k < \ell \leqslant n$$

$$= \Sigma\Sigma \; b_{ij} \log \left(x_{(j)} - x_{(i)}\right) \quad , \quad 1 \leqslant i < j \leqslant n$$

เมื่อ $b_{ij} = \binom{n}{4}^{-1} \left\{ 2(n-j)(i-1) - \binom{n-j}{2} - \binom{i-1}{2} \right\}$

T_2' มีความไวต่อความโด่ง

การคำนวณหาค่าวิกฤติของตัวสถิติทั้งสอง ประมาณด้วยการแจกแจงปกติ คือ

$$z_1 = \frac{T_1' - E(T_1')}{\sqrt{v(T_1')}}$$

และ $$z_2 = \frac{T_2' - E(T_2')}{\sqrt{v(T_2')}}$$

โดยที่ค่าเฉลี่ยและความแปรปรวนของตัวสถิติ จะประมาณด้วยเทคนิคมอนติคาร์โลสุ่มเลขขึ้น จาก 2,000 ตัวอย่าง เมื่อขนาดตัวอย่างต่าง ๆ กัน ซึ่งจะแสดงผลในตารางที่ 5 ในภาคผนวก

ตัวอย่างที่ 5 การใช้ตัวสถิติของ Hannu Oja ในการทดสอบการแจกแจงปกติ

ข้อมูลตัวอย่าง	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
	-4	-2	0	1	5	6	8

H_o : ประชากรมีการแจกแจงแบบปกติ

H_1 : ประชากรไม่มีการแจกแจงแบบปกติ

i	j	$x_{(j)}-x_{(i)}$	$\log(x_{j})-x_{(i)}$	$a_{ij}\ \log(x_{(j)}-x_{(i)})$	$b_{ij}\ \log\ (x_{(j)}-x_{(i)})$
1	2	2	0.30103	-0.04300	-0.08601
1	3	4	0.60206	-0.06881	-0.10321
1	4	5	0.69897	-0.05991	-0.05991
1	5	9	0.95424	-0.05453	-0.02726
1	6	10	1.00000	-0.02857	0
1	7	12	1.07180	0	0
2	3	2	0.30103	-0.02580	0.01720
2	4	3	0.47712	-0.02727	0.04090
2	5	7	0.84510	-0.02414	0.07244
2	6	8	0.90390	0	0.05161
2	7	10	1.0000	0.02857	0
3	4	1	0	0	0
3	5	5	0.69897	0	0.11982
3	6	6	0.77815	0.02223	0.06670
3	7	8	0.90309	0.0516	0.02580
4	5	4	0.60206	0.01721	0.13761
4	6	5	0.69897	0.03994	0.03994
4	7	7	0.84510	0.07243	0
5	6	1	0	0	0
5	7	3	0.47712	0.05453	-0.02726
6	7	2	0.30103	0.04301	-0.04300
				-0.00251	0.22537

$$T_1' = \sum_{i=1}^{7} \sum_{i=1}^{6} a_{ij} \log (x_{(j)} - x_{(i)})$$

$$a_{ij} = \binom{n}{3}^{-1} (i+j-n-1) \qquad 1 \leqslant i < j \leqslant 7$$

$$\therefore \quad T_1' = -0.00251$$

โดยการกะประมาณจากการแจกแจงปกติ จากตารางที่ 5 ในภาคผนวก

$$E(T_1') = -0.00983$$

$$S(T_1') = \sqrt{0.05096} = 0.22574$$

$$\text{และ} \quad z = \frac{-0.00251 - (-0.00983)}{0.22574} = 0.03243$$

สมมติว่ากำหนดให้ระดับนัยสำคัญเท่ากับ 0.01 จุดวิกฤติของค่า z คือ $\pm$ 2.57 เนื่องจากค่าของ z ที่คำนวณได้เท่ากับ 0.03243 เราจึงยอมรับ H_o ที่ตั้งไว้ และสรุปว่าประชากรมีการแจกแจงปกติ ในทำนองเดียวกันจะสามารถคำนวณ T_2' ได้ดังนี้

$$T_2' = \sum_{j=2}^{7} \sum_{i=1}^{6} b_{ij} \log (x_{(j)} - x_{(i)})$$

$$\text{เมื่อ} \quad b_{ij} = \binom{n}{4}^{-1} \left\{ 2(n-j)(i-1) - \binom{n-j}{2} - \binom{i-1}{2} \right\} \quad \text{และ} \quad 1 \leqslant i < j \leqslant 7$$

$$\therefore \quad T_2' = 0.22537$$

จากตารางที่ 5 $E(T_2') = 0.11681$ และ $S(T_2') = 0.36551$

$$\therefore \quad z = \frac{0.22537 - 0.11681}{0.35551} = 0.30536$$

ที่ระดับนัยสำคัญ 0.01 จุดวิกฤติของค่า z คือ $\pm$ 2.57 ดังนั้นจะยอมรับ H_o และสรุปว่าประชากรมีการแจกแจงปกติ

2.2 ตัวสถิติอื่น ๆ บางตัว สำหรับการทดสอบการแจกแจงปกติ

เนื่องจากตัวสถิติสำหรับการทดสอบการแจกแจงปกติมีจำนวนมาก ในส่วนนี้จึงจะเสนอตัวสถิติเพียงบางตัวเท่านั้น โดยจะเสนอตัวสถิติที่เป็นที่นิยมใช้ในการทดสอบหรือตัวสถิติที่มีความสัมพันธ์กับตัวสถิติที่ศึกษาในงานวิจัยนี้

2.2.1 ทดสอบความเบ้ (Test of Skewness)

$$\sqrt{b_1} = g_1 = \frac{m_3}{(\sqrt{m_2^{\,3}})}$$

เมื่อ $$m_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{n}$$

ภายใต้การแจกแจงปกติ g_1 จะเข้าใกล้ศูนย์ ในการตัดสินใจว่า g_1 แตกต่างจากศูนย์อย่างมีนัยสำคัญหรือไม่ ใช้การแจกแจงปกติประมาณการแจกแจงของ g_1 โดยมีค่าเฉลี่ย 0 และส่วนเบี่ยงเบนมาตรฐานเป็น $\sqrt{b/n}$ การทดสอบนี้จะถูกต้อง เมื่อตัวอย่างมีขนาด (n) ตั้งแต่ 150 ขึ้นไป สำหรับ n = 25 ถึง 200 ค่า g_1 ซึ่งมีนัยสำคัญที่ระดับ 0.05 และ 0.01 ในการทดสอบทางเดียว ใช้ตารางที่ 6 ในภาคผนวก

2.2.2 ทดสอบความโด่ง (Test for Kutosis)

$$g_2 = b_2 - 3 = \frac{m_4}{(m_2)^2} - 3$$

เมื่อ $$m_k = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^k}{n}$$

ค่าวิกฤติใช้การแจกแจงปกติประมาณการแจกแจงของ g_1 โดยมีค่าเฉลี่ย 0 และส่วนเบี่ยงเบนมาตรฐานเป็น $\sqrt{24/n}$ เมื่อขนาดตัวอย่างมากกว่า 1,000 สำหรับ $n = 50$ ถึง 1,000 ค่า g_2 ที่ระดับนัยสำคัญ 0.05 และ 0.01 ใช้ตารางที่ 7 ในภาคผนวก

2.2.3 D' Agostino Statistic (D) เป็นตัวสถิติทดสอบที่ D' Agostino (1971 : 341-348) ได้สร้างขึ้นโดยเปลี่ยนแปลงจากตัวสถิติของ Shapiro และ Wilk

$$D = \frac{T}{n^2 S}$$

เมื่อ $$T = \sum_{i=1}^{n} \{ i - \frac{1}{2}(n-1)\} x_{(i)}$$

ภายใต้การแจกแจงปกติ

$$E(D) = \frac{(n-1)\tau\{\frac{n-1}{2}\}}{2(2n\Pi)^{\frac{1}{2}}\tau\left(\frac{n}{2}\right)} \simeq \frac{1}{2\sqrt{\Pi}}$$

$$SD.(D) = \left\{\frac{12\sqrt{3} - 37 + 2\Pi}{24\Pi n}\right\}^{\frac{1}{2}} = \frac{0.02998598}{\sqrt{n}}$$

ดังนั้น $$y = \frac{\{D - (2\sqrt{\Pi})^{-1}\}}{S.D.} \sim N(0,1)$$

จากการศึกษาโดยการซิมูเลชั่น ถ้าการแจกแจงมีความโด่งต่ำกว่า 3 ค่าของ y จะมากกว่า 0 และถ้าการแจกแจงมีความโด่งมากกว่า 3 ค่าของ y จะน้อยกว่า 0

2.2.4 Shapiro-Francia Statistic (w') (Shapiro และ Francia 1972 : 215-216) ใช้สำหรับตัวอย่างขนาดใหญ่ และค่าสังเกต (cbserve) มีความเป็นอิสระ (independent)

$$w' = \frac{(\underset{\sim}{b}'\underset{\sim}{x})^2}{nS^2}$$

เมื่อ $$\underset{\sim}{b}' = \frac{\underset{\sim}{c}'}{(\underset{\sim}{c}'\underset{\sim}{c})^{\frac{1}{2}}}$$

c_i เป็นค่าคาดหวังของ Order Statistic ลำดับที่ i จากประชากรที่มีการแจกแจงปกติ ค่าของ $\underset{\sim}{c}$ สามารถเปิดจากตารางของ Harter (1961:151-165)

2.2.5 การทดสอบโดยใช้ Empirical Distribution Funtion

ให้ ϕ คือการแจกแจงปกติ N(0 , 1)

$$z_i = \phi \left\{ \frac{x_{(i)} - \mu}{\sigma} \right\} \qquad i = 1, 2, \ldots, n$$

1) Kolmogorov Smirnov Statistic (K)

$$K = \max (D^+ , D^-)$$

$$D^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - z_i \right\}$$

$$D^- = \max_{1 \leq i \leq n} \left\{ z_i - \frac{(i-1)}{n} \right\}$$

2) Cramer-Von Mises Statistic (w^2)

$$w^2 = \sum_{i=1}^{n} \left\{ z_i - \frac{(2i - 1)}{2n} \right\}^2 + \left(\frac{1}{12n} \right)$$

3) The Watson Statistic (u^2)

$$u^2 = w^2 - n \{\bar{z} - \frac{1}{2} \}^2$$

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

4) The Anderson-Darling Statistic (A^2)

$$A^2 = - \sum_{i=1}^{n} \left[(2i - 1) \{\log z_i + \log (1-z_{n+1-i}) \}/_n \right] -n$$

ค่าวิกฤติของตัวสถิติ K เมื่อขนาดตัวอย่างน้อยกว่า 100 ดูได้จากตารางของ Birnbaum (1952 : 225-441) และค่าวิกฤตของตัวสถิติ w^2 , u^2 และ A^2 ดูได้จากตารางที่ 2 ของ Durbin, Knott และ Taylor (1975:1961)

2.3 ผลงานวิจัยที่เกี่ยวข้อง

สำหรับงานวิจัยที่เกี่ยวข้องกัน ได้มีนักสถิติจำนวนมากได้ศึกษาอำนาจการทดสอบของตัวสถิติที่ใช้ในการทดสอบการแจกแจงปกติ ซึ่งในส่วนนี้จะเสนอเฉพาะบางผลงานวิจัยเท่านั้น

Shapiro และคณะ (1968 : 1343-1372) เป็นผู้เริ่มในการศึกษาอำนาจการทดสอบของตัวสถิติต่าง ๆ ในการทดสอบการแจกแจงปกติ โดยทำการศึกษาตัวสถิติ 9 ตัวคือ Shapiro-Wilk Statistic (W) , $\sqrt{b_1}$, b_2 , Kolmogorov Smirnov (K), Cramer-Von Mises (w^2), Anderson-Darling (A^2) , Durein (T), Chi-Sqaure Test (χ^2) และ Studentized Range Test (u) ภายใต้การแจกแจง 12 การแจกแจง ซึ่งมีค่าพารามิเตอร์ที่แตกต่างกัน รวมเป็น 45 การแจกแจงได้ผลสรุปดังนี้

1. Shapiro-Wilk Statistic ใช้ได้ดีในการทดสอบทั่วไป

2. การทดสอบโดยใช้ Empirical Distribution Funtion จะได้อำนาจการทดสอบต่ำ

3. Studentized Range Test จะมีอำนาจการทดสอบสูงเมื่อประชากรมีการแจกแจงแบบสมมาตรและหางสั้น และจะมีอำนาจการทดสอบต่ำเมื่อประชากรมีการแจกแจงแบบเบ้

4. $\sqrt{b_1}$ และ b_2 ใช้ในการทดสอบได้ดี แต่มีอำนาจการทดสอบต่ำกว่า W