บทที่ 2

ตัวสถิติและผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงรายละเอียดเกี่ยวกับความรู้พื้นฐานของชนิดของข้อมูลขาดหาย (type of censoring) และอัตราการเสีย (hazard rate) ของข้อมูลรวมทั้งตัวสถิติที่ใช้ทดสอบการแจกแจงแบบเอกซ์โปเนนเชียล พร้อมทั้งตัวอย่างการใช้ตัวสถิติและผลงานวิจัยที่เกี่ยวข้อง ซึ่งรายละเอียดต่าง ๆ มีดังนี้

2.1 <u>ความรู้พื้นฐาน</u>

2.1.1 <u>ชนิดของข้อมูลขาดหาย</u> (type of censoring)

ก. ข้อมูลขาดหายประเภทที่ 1 (type 1 censoring) เรียกอีกอย่างหนึ่งว่า time censoring ข้อมูลขาดหายประเภทนี้จะกำหนดเวลาของการเกิดข้อมูลขาดหายไว้ล่วงหน้าแล้ว เช่น การศึกษาอายุการใช้งานของเครื่องจักรจำนวนหนึ่ง กำหนดอายุใช้งานของเครื่องจักร คือ 20,000 ชั่วโมง ถ้าเครื่องจักรทำงานเกิน 20,000 ชั่วโมงถือว่าอายุการใช้งานของเครื่องจักรเครื่องที่เกินเป็นข้อมูลขาดหาย (censored data) เนื่องจากไม่สามารถบันทึกข้อมูลได้ กรณีที่เครื่องจักรบางเครื่องมีอายุการใช้งานยาวนานและชั่วโมงที่กำหนดถือว่าเป็น time censoring

ข. ข้อมูลขาดหายประเภทที่ 2 (type 2 censoring) ในบางกรณีผู้วิจัยไม่สามารถกำหนด time censoring ที่เหมาะสมได้ ดังนั้นจึงกำหนดข้อมูลที่ไม่ขาดหายแทน โดยจะหยุดการทดลองเมื่อข้อมูลที่ไม่ขาดหายเกิดขึ้นครบตามจำนวนที่กำหนดไว้

ค. ข้อมูลขาดหายแบบสุ่ม (random censoring) ข้อมูลขาดหายประเภทนี้ ส่วนใหญ่พบในข้อมูลที่เกี่ยวกับการแพทย์ หลังจากได้รับการรักษาแล้วซึ่งจะทำให้ไม่สามารถได้ข้อมูลที่แน่นอน

2.1.2 <u>อัตราการเสีย</u> (hazard rate หรือ hazard function)-

ให้ T เป็นตัวแปรสุ่มซึ่งแทนอายุใช้งาน (time to failure)

$f(t)$ แทน ฟังชันความหนาแน่น

$F(t)$ แทนฟังชันการแจกแจงสะสม

และ $R(t)$ (reiability function) คือความน่าจะเป็นที่หน่วยตัวอย่างมีอายุใช้งานได้นานกว่าเวลา t

ดังนั้น $R(t) = p(T>t)$

$= 1 - F(t)$

$h(t)$ แทนอัตราการเสีย หรือฟังก์ชันการเสียมีค่าเท่ากับค่าจำกัด (limit) ของความน่าจะเป็นที่แต่ละหน่วยตัวอย่างเสียในช่วงเวลาสั้น ๆ จาก t ถึง $t+\Delta t$ ต่อหน่วยเวลา Δt เมื่อแต่ละหน่วยตัวอย่างมีอายุใช้งานมากกว่า t นั่นคือ

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t<T<t+\Delta t \mid T>t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{(1-F(t)).\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{R(t) - R(t+\Delta t)}{R(t).\Delta t}$$

$$= -\frac{R'(t)}{R(t)}$$

$$= \frac{f(t)}{R(t)}$$

$$= \frac{f(t)}{1-F(t)}$$

โดย $h(t)$ มีคุณสมบัติดังนี้

ก. $h(t) > 0$ เมื่อ $-\infty < t < \infty$

ข. $\lim_{t \to -\infty} \int_{-\infty}^{t} h(t)dt = 0$ และ $\lim_{t \to \infty} \int_{-\infty}^{t} h(t)dt = \infty$

โดยทั่วไปฟังก์ชันการเสียมีรูปแบบเป็นฟังก์ชันลด ฟังก์ชันเพิ่ม และฟังก์ชันคงที่ ดังแสดงไว้ในบทที่ 1 และเมื่อทราบฟังก์ชันการเสียของข้อมูลชุดหนึ่ง สามารถบอกได้ว่าข้อมูลมีฟังก์ชันความหนาแน่น $f(t)$ และฟังก์ชันการแจกแจงสะสม $F(t)$ โดยใช้คุณสมบัติต่อไปนี้

จาก $h(t) = \frac{f(t)}{1-F(t)}$

$$= \frac{-R'(t)}{R(t)}$$

$$= \frac{-1}{R(t)} \frac{d\,R(t)}{d(t)}$$

$$h(x) = \frac{-1}{R(x)} \frac{d\,R(x)}{d(x)}$$

$$-\int_0^t h(x)d(x) = \int_0^t \frac{1}{R(x)} d\,R(x)$$

$$= \ln R(t) - \ln (R(0))$$

เพราะว่า $F(0) = 0$ ดังนั้น $R(0) = 1- F(0) = 1$

$$-\int_0^t h(x)d(x) = \ln R(t) \qquad (\ln R(0) = \ln(1) = 0)$$

ดังนั้น $$R(t) = \exp\left(-\int_0^t h(x)d(x) \right)$$

นั่นคือ $$F(t) = 1 - \exp\left(-\int_0^t h(x)d(x) \right)$$

นั่นคือเมื่อทราบว่าฟังก์ชันการเสียเป็นค่าคงที่เท่ากับ $1/\beta$ สามารถหาฟังก์ชันการแจกแจงสะสมได้ดังนี้

$$F(t) = 1 - \exp\left(-\int_0^t (1/\beta)d(x)\right)$$

$$F(t) = 1 - \exp(-t/\beta) \quad , t>0$$

ซึ่งก็คือ ฟังก์ชันการแจกแจงสะสมของการแจกแจงแบบเอกซ์โปเนนเชียล

2.2 ตัวสถิติทดสอบ

2.2.1 ตัวสถิติทดสอบ Z (Regression Test)

Brain และ Shapriro (ค.ศ 1983) ได้เสนอตัวสถิติทดสอบการแจกแจง

แบบเอกซ์โปเนนเชียล โดยตัวสถิติทดสอบ Z เป็นฟังก์ชันค่าความชันของเส้นถดถอยซึ่งเป็นความสัมพันธ์ของตัวแปรตาม y_i กับตัวแปรอิสระ i กล่าวคือ ตัวสถิติทดสอบอยู่ในรูปของ

$$Z = \frac{[12 / (m-2)]^{1/2} \sum_{i=1}^{m-1} a_i Y_{r1+i+1}}{\sum_{i=1}^{m-1} Y_{r1+i+1}}$$

เมื่อ $x_1 < x_2 < \dots < x_n$ เป็นข้อมูลอันดับขนาด n

กำหนด $Y_i = (n-i+1)(x_i - x_{i-1})$, i=2,3,..., n

ซึ่ง Pyke (ค.ศ 1965) สามารถแสดงได้ว่าในกรณีตัวอย่างขนาดใหญ่ Y_i เป็นตัวแปรสุ่มแบบเอกซ์โปเนนเชียลและเป็นอิสระต่อกัน

$a_i = i - m/2$,i= 1,2,...,m-1

$m = n-r_1-r_2$

x_{r1+1} = ค่าสังเกตที่เล็กที่สุด

x_{n-r2} = ค่าสังเกตที่ใหญ่ที่สุด

เมื่อ r_1 เป็นจำนวนข้อมูลขาดหายทางซ้าย (left censored data)

r_2 เป็นจำนวนข้อมูลขาดหายทางขวา (right censored data)

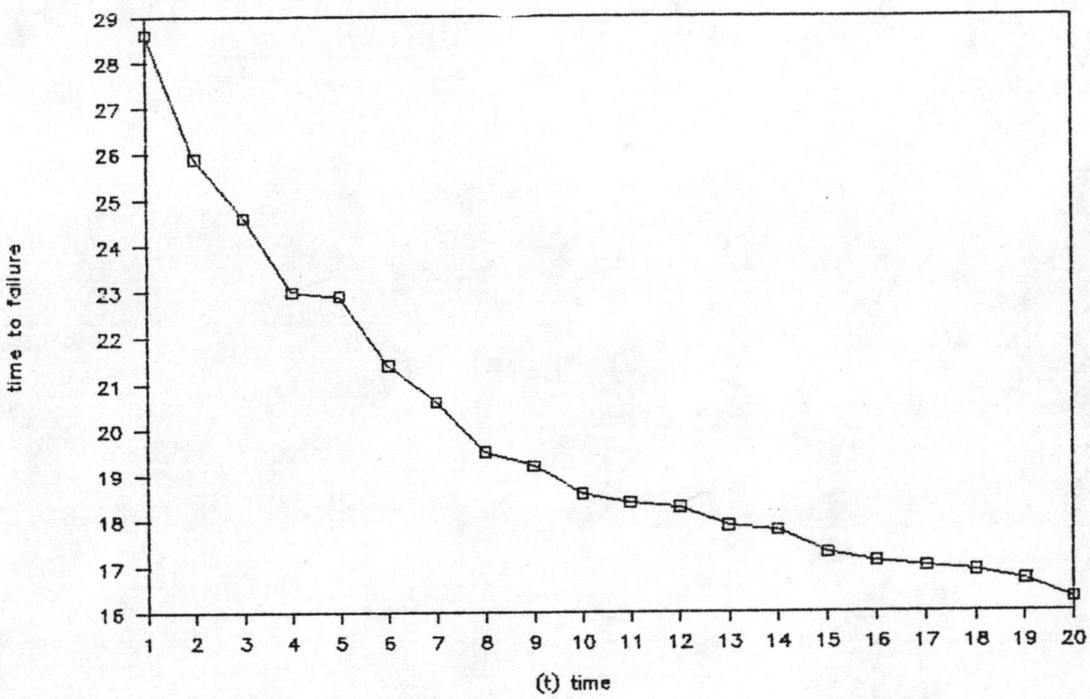
เราจะปฏิเสธสมมุติฐาน H_0 : ข้อมูลมีการแจกแจงแบบเอกซ์โปเนนเชียล

เมื่อ $|Z| > Z^*_{\alpha/2}$, $Z^*_{\alpha/2} \sim N(0,1)$

<u>ตัวอย่างที่ 1</u> การทดสอบอายุการใช้งานของเครื่องจักร 35 เครื่อง บันทึกข้อมูลโดยคิดเวลาในการทำงานเป็นชั่วโมงกำหนดการทดสอบให้หยุดภายหลัง 30 ชั่วโมงของการทดสอบปรากฏว่ามีเครื่องจักร 4 เครื่องที่หยุดทำงาน ก่อนที่จะบันทึกข้อมูล และหลังจาก 30 ชั่วโมงมีเครื่องจักร 10 เครื่องที่ยังทำงานต่อซึ่งไม่สามารถเก็บข้อมูลได้ สำหรับข้อมูลที่เก็บได้เป็นดังนี้

-,-,-,-,16.3,16.7,16.9,17.0,17.1,17.3,17.8,17.9,18.3,18.4,18.6, 19.1,19.5,20.6,21.4,22.9,23.0,24.6,25.9,28.6,-,-,-,-,-,-,-,-,-,-

ซึ่งสามารถเขียนเป็นรูปกราฟได้ดังนี้



รูปที่ 2.1

แสดงกราฟของข้อมูลได้จากตัวอย่างที่ 1

กำหนด H_0 : ข้อมูลมีการแจกแจงแบบเอกซ์โปเนนเชียล

H_1 : ข้อมูลไม่มีการแจกแจงแบบเอกซ์โปเนนเชียล

จากตัวอย่างนี้ $n = 35$, $r_1 = 4$, $r_2 = 10$, $m = 35-10-4 = 21$

จาก $Y_i = (n-i+1)(x_i-x_{i-1})$

สามารถจัดเรียงข้อมูลใหม่ได้ดังนี้

i	x_t	y_t	a_t=i-m/2	a_t*y_{5+t}
5	16.0	-	-	-
6	16.3	9.0	-9.5	-85.5
7	16.7	11.6	-8.5	-98.5
8	16.9	5.6	-7.5	-42.00
9	17.0	2.7	-6.5	-17.55
10	17.1	2.6	-5.5	-14.30
11	17.3	5.0	-4.5	-22.50
12	17.8	12.0	-3.5	-42.00
13	17.9	2.3	-2.5	-5.75
14	18.3	8.8	-1.5	-13.20
15	18.4	2.1	-0.5	-1.05
16	18.6	4.0	0.5	2.00
17	19.1	9.5	1.5	14.25
18	19.5	7.2	2.5	18.00
19	20.6	18.7	3.5	65.45
20	21.4	12.8	4.5	57.60
21	22.9	22.5	5.5	123.75
22	23.0	1.4	6.5	9.10
23	24.6	20.8	7.5	156.00
24	25.9	15.6	8.5	132.60
25	28.6	29.7	9.5	282.15
		รวม 203.9		รวม 518.45

จากตารางจะได้ว่า $\sum_{i=1}^{20} a_i Y_{s+i} = 518.45$ และ $\sum_{i=1}^{20} a_i Y_{s+i} = 203.9$

$$Z = \frac{[\ 12\ /\ 21-2\ ]^{1/2}\ .\ 518.45}{203.9}$$

$$= 2.021$$

จากตารางการแจกแจงปกติมาตรฐาน เมื่อกำหนดระดับนัยสำคัญ 0.05

จะได้ว่า $Z^*_{\alpha/2} = 1.96$

ดังนั้น $Z > Z^*_{\alpha/2}$

เราจะปฏิเสธสมมุติฐาน H_0 ณ ระดับนัยสำคัญ 0.05

2.2.2 <u>ตัวสถิติทดสอบ Gnedenko F Test</u>

Gnedenko และคณะ (ค.ศ1969)ได้เสนอตัวสถิติทดสอบ F เพื่อทดสอบการแจกแจงแบบเอกซ์โปเนนเชียล โดยตัวสถิติทดสอบ F หาได้จากอัตราส่วนของผลรวมเฉลี่ยของข้อมูล 2 ชุด ชุดแรก (s1) มี k ข้อมูล ชุดที่ 2 (s2) มี m-1-k ข้อมูล ตัวสถิติทดสอบคือ

$$F = s1\ /\ s2$$

โดยที่ $s1 = \sum_{i=2}^{k+1} Y_{r1+i}\ /\ k$

$$s2 = \sum_{i=k+2}^{m} Y_{r1+i}\ /\ (m-1-k)$$

$$m = n-r_1-r_2$$

และ k = จำนวนเต็มใด ๆ = $(m-1)\ /\ 2$

เราจะปฏิเสธสมมุติฐาน H_0 เมื่อ $F > F^*_{\alpha/2}\,[2k, 2(m-1-k)]$

หรือ $1/F > F^*_{\alpha/2}\,[2(m-1-k), 2k]$

โดยที่ $F^*_{\alpha/2}\,(m,n)$ คือค่าในตาราง F ณ ระดับความเป็นอิสระ m และ n

<u>ตัวอย่างที่ 2</u> จากข้อมูลในตัวอย่างที่ 1 ได้ $m=21, n=35, r_1=4, r_2=10, k=(21-1)/2=10$

กำหนดให้ H_0 : ข้อมูลมีการแจกแจงแบบเอกซ์โปเนนเชียล

H_1 : ข้อมูลไม่มีการแจกแจงแบบเอกซ์โปเนนเชียล

พิจารณาค่า s_1 และ s_2

$$s_1 = \sum_{i=2}^{11} Y_{4+i} / 10$$

$$= (9+11.6+5.6+2.7+2.6+5.0+12.0+8.3+8.8+2.1)/10$$

$$= 6.17$$

$$s_2 = \sum_{i=10+2}^{21} Y_{4+i} / (21-1-10)$$

$$= (4.0+9.5+7.2+18.7+12.8+22.5+1.4+20.8+15.6+29.7)/10$$

$$= 14.22$$

ดังนั้นได้ค่า $F = s1 / s2$

$= 6.17 / 14.22$

$= 0.4338$

จากตาราง F ได้ $F^*_{0.05}\,[\,2(10), 2(21-1-10)\,] = F^*_{0.05}\,(20,20) = 2.12$

ดังนั้น $1/F = 2.3047 > F^*_{0.05} = 2.12$

เราจึงปฏิเสธสมมุติฐาน H_0 ณ ระดับนัยสำคัญ 0.10

2.2.3 ตัวสถิติทดสอบ K (Kolmogorov-Smirnov Test)

Stephen(ค.ศ 1974) ได้เสนอตัวสถิติทดสอบ K เพื่อทดสอบการแจกแจงแบบเอกซ์โปเนนเชียล โดยพิจารณาจากค่าแตกต่างที่สูงที่สุดระหว่างฟังก์ชันการแจกแจงสะสมของตัวอย่างที่จะวิเคราะห์กับฟังก์ชันการแจกแจงสะสมของการแจกแจงที่ต้องการเปรียบเทียบ ตัวสถิติทดสอบคือ

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

โดยที่ $D^+ = \max(i/n, F(x_i))$

$D^- = \max(F(x_i), (i-1)/n)$

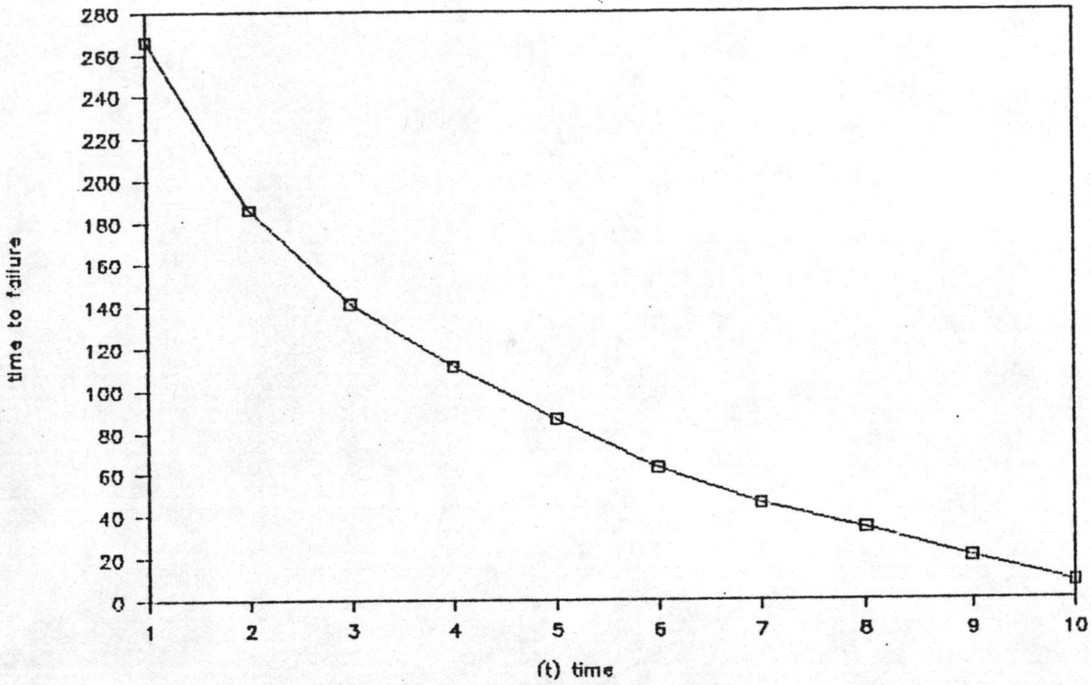
และ $F(x_i)$ คือฟังก์ชันการแจกแจงสะสม , $(i=1,2,\ldots,n)$ และ n=จำนวนข้อมูล

เราจะปฏิเสธสมมุติฐาน H_o เมื่อ $K > K^*_{\alpha/2,n}$ (จากตาราง)

ตัวอย่างที่ 3 ถ้าอายุการใช้งานของเครื่องจักรชุดหนึ่งเป็นดังนี้

8,20,34,46,63,86,111,141,186,266

ซึ่งสามารถแสดงเป็นรูปกราฟได้ดังนี้



รูปที่ 2.20

แสดงกราฟของข้อมูลจากตัวอย่างที่ 3

กำหนด H_0 : ข้อมูลมีการแจกแจงแบบเอกซ์โปเนนเชียล

H_1 : ข้อมูลไม่มีการแจกแจงแบบเอกซ์โปเนนเชียล

จากข้อมูลที่กำหนดให้เราสามารถจัดเรียงข้อมูลใหม่ได้ตามตารางต่อไปนี้

i	x_i	(i-1)/n	i/n	$F(x_i)$	i/n - $F(x_i)$	$F(x_i)$ - (i-1)/n
1	8	0.00	0.10	0.077	0.023	0.077
2	20	0.10	0.20	0.181	0.019	0.081
3	34	0.20	0.30	0.288	0.012	0.088
4	46	0.30	0.40	0.369	0.031	0.064
5	63	0.40	0.50	0.467	0.033	0.067
6	86	0.50	0.60	0.577	0.023	0.072
7	111	0.60	0.70	0.670	0.030	0.070
8	141	0.70	0.80	0.756	0.044	0.056
9	186	0.80	0.90	0.844	0.056	0.044
10	266	0.90	1.00	0.930	0.070	0.030

จากตาราง $F(x_i) = 1 - \exp(-x_i / x)$

เมื่อ $x = \sum_{i=1}^{n} x_i / n$

เพราะฉะนั้น $D^{+} = \max (i/n , F(x_i))$

$= 0.070$

$D^{-} = \max (F(x_i) , (i-1)/n)$

$= 0.088$

ดังนั้น $K = \max (D^{+}, D^{-})$

$= 0.088$

พิจารณาค่า $K^{*}_{\alpha/2}$ จากตาราง Kolmogorov-Smirnov

ได้ $K^{*}_{(0.05,10)} = 0.3244$

ดังนั้น $K = 0.088 < K^{*}_{(0.05,10)} = 0.3244$

จึงไม่เหตุผลที่จะปฏิเสธสมมุติฐาน H_o กล่าวคือ ข้อมูลชุดนี้มีการแจกแจงแบบเอกซ์โปเนนเชียล

จากตัวอย่างเดียวกันนี้ เมื่อทดสอบโดยใช้ตัวสถิติทดสอบ Z สามารถแสดงได้จากตารางต่อไปนี้

i	x_i	$Y_i = (n - i + 1)(X_i - X_{i-1})$	$a_i = i - m/2$	$a_i \cdot Y_i$
1	8			
2	20	108	− 3	− 324
3	34	112	− 2	− 224
4	46	84	− 1	− 84
5	63	102	0	0
6	86	115	1	115
7	111	100	2	200
8	141	90	3	270
9	186	90	4	360
10	266	80	5	400
		รวม 881		รวม 713

จาก
$$Z = \frac{[12 / (m-2)]^{1/2} \sum_{i=1}^{m-1} a_i Y_{ri+i+1}}{\sum_{i=1}^{m-1} Y_{ri+i+1}}$$

$$= \frac{[12 / (10-2)]^{1/2} \times 713}{881}$$

$$= 0.991$$

ดังนั้น $Z = 0.991 < Z^{*}_{\alpha/2} = 1.64 \quad (\alpha = 0.10)$

ดังนั้นจึงไม่มีเหตุผลที่จะปฏิเสธสมมุติฐาน H_o เช่นเดียวกันกับตัวสถิติทดสอบ K

2.3 ผลงานวิจัยที่เกี่ยวข้อง

ผลงานวิจัยของนักสถิติที่สนใจทดสอบการแจกแจงแบบเอกซ์โปเนนเชียลมีมากมาย ในที่นี้จะเสนอผลงานวิจัยบางผลงานวิจัยเท่านั้น

Danny Dyer และ Mickie Sui Habin (ค.ศ 1981:278-291) ได้ศึกษาอำนาจการทดสอบของตัวสถิติทดสอบ 3 ตัวที่ใช้ทดสอบการแจกแจงแบบเอกซ์โปเนนเชียล คือ Tiku Test , Shapiro-Wilk Test และ Durbin Test ภายใต้ประชากรศึกษา 16 ประชากร โดยศึกษาเฉพาะขนาดตัวอย่างขนาดเล็ก คือ n =5,10,15 และ 25 ได้ผลสรุปดังนี้

1) โดยทั่วไป Tiku Test มีอำนาจการทดสอบสูงกว่าตัวสถิติทดสอบตัวอื่น

2) เมื่อขนาดตัวอย่างขนาดเล็ก (n=5,10) พบว่าตัวสถิติทดสอบ Shapiro-Wilk Test มีอำนาจการทดสอบสูงในบางกรณีของการแจกแจงแย้ง[1] (alternative distributions)

3) โดยเฉลี่ยตัวสถิติทดสอบ Dubin Test ให้อำนาจการทดสอบต่ำ เมื่อเทียบกับตัวสถิติทดสอบ Tiku Test และ Shapiro-Wilk Test

---

1 การแจกแจงแย้ง ในที่นี้หมายถึง การแจกแจงของข้อมูลที่กำหนดให้มีการแจกแจงเป็นแบบอื่น ๆ ที่ไม่ใช่การแจกแจงแบบเอกซ์โปเนนเชียล