บทที่ 3
ความรู้พื้นฐานเกี่ยวกับสถิติที่ใช้วิเคราะห์

ในบทนี้จะกล่าวถึงความรู้พื้นฐานเกี่ยวกับสถิติที่ใช้วิเคราะห์ในงานวิจัยนี้ คือ

3.1 วิธีการวิเคราะห์แบบโปรไฟล์ (Profile Analysis)

3.2 การวิเคราะห์จำแนกประเภท (Discriminant Analysis)

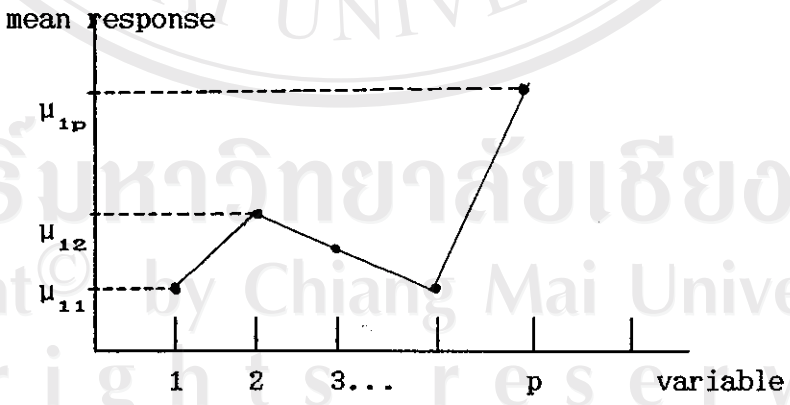
3.3 การหาค่าวินิจฉัย (Diagnostic value)

3.1 วิธีการวิเคราะห์แบบโปรไฟล์ (Profile Analysis)

วิธีการวิเคราะห์แบบโปรไฟล์ เป็นวิธีการที่ใช้ในการวิเคราะห์ว่าจะสามารถรวมกลุ่มประชากร 2 กลุ่มที่เป็นอิสระต่อกัน เข้าเป็นกลุ่มประชากรเดียวกันได้หรือไม่ โดยแต่ละกลุ่มประชากรมีตัวแปรจากการทดลอง p ตัวแปร โดยที่ค่าเฉลี่ยประชากร (population means) ของประชากรกลุ่มที่ 1 จะนิยามได้ดังนี้

$$\mu'_1 = [\mu_{11}, \mu_{12}, \ldots \mu_{1p}]$$

โดยที่ $\mu_{11}, \mu_{12}, \ldots \mu_{1p}$ เป็นค่าเฉลี่ยของ p ตัวแปร จากประชากรกลุ่มที่ 1 และแสดงเป็นแผนภาพโปรไฟล์ดังนี้



<u>แผนภาพที่ 3.1</u> เส้นโปรไฟล์ของประชากรกลุ่มที่ 1

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ค่าเฉลี่ยประชากรของประชากรกลุ่มที่ 2 จะเขียนได้ในลักษณะเดียวกัน คือ

$$\mu'_2 = [\mu_{21}, \mu_{22}, \ldots \mu_{2p}]$$

โดยที่

$\mu_{21}, \mu_{22}, \ldots \mu_{2p}$ เป็นค่าเฉลี่ยของ p ตัวแปร จากประชากรกลุ่มที่ 2 ทำการวิเคราะห์ว่าจะสามารถรวมประชากร 2 กลุ่มที่มี p ตัวแปร เข้าเป็นกลุ่มเดียวกันได้หรือไม่ โดยตั้งสมมติฐานว่าการทดลองแต่ละตัวแปรมีผล (effect) แบบเดียวกันในประชากรทั้ง 2 กลุ่ม หรือ

$$H_o : \mu_1 = \mu_2$$

ในรูปแบบของโปรไฟล์ มีวิธีการวิธีการวิเคราะห์ตามขั้นตอน ดังนี้คือ

1. การทดสอบการขนานกัน (Test of Parallelism) เป็นการทดสอบว่าเส้นโปรไฟล์ขนานกันหรือไม่

- ถ้าขนานกัน จะทดสอบต่อไปในขั้นที่ 2 และ 3
- ถ้าไม่ขนานกัน แสดงว่าไม่สามารถรวมประชากร 2 กลุ่มนี้ เป็นกลุ่มเดียวกันได้

2. การทดสอบการทับกัน (Test of Coincident) จะทดสอบว่าเส้นโปรไฟล์ที่ขนานกันนั้นจะทับกันสนิท (ระดับของตัวแปรจะเท่ากัน) หรือไม่

3. การทดสอบการอยู่ในแนวระดับเดียวกับค่าเฉลี่ย (Test for Equal response means) เป็นการทดสอบว่าเส้นโปร์ไฟล์ที่ขนานนั้น จะมีค่าเฉลี่ยของโปรไฟล์อยู่ในแนวระดับเดียวกันหรือไม่

1. Test of Parallelism เป็นการทดสอบดูว่าเส้นโปรไฟล์ ขนานกันหรือมีระยะห่างเท่ากันหรือไม่ วิธีการทดสอบมี ดังนี้

สมมติฐาน $H_o: C\underset{\sim}{\mu}_1 = C\underset{\sim}{\mu}_2$

โดยที่

C = (p-1) × p transformative matrix

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}_{(p-1) \times p}$$

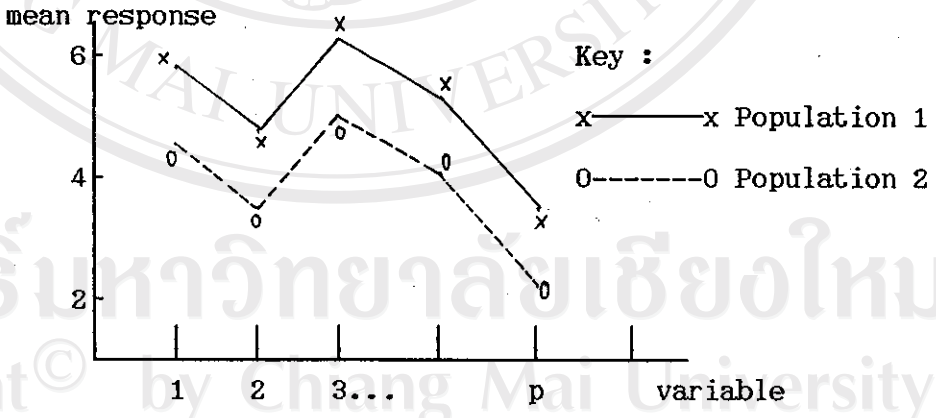
ค่าสถิติที่ใช้ $T^2 = \frac{n_1 n_2}{n_1+n_2} [C(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)]' [CSC']^{-1} [C(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)]$

$$F_c = \frac{(n_1+n_2-p)}{(n_1+n_2-2)(p-1)} \cdot T^2$$

และจะยอมรับสมมติฐาน H_o (นั่นคือ เส้นโปรไฟล์ขนานกัน) เมื่อ

$$F_c < F_{\alpha\ ;\ (p-1),(n_1+n_2-p)}$$

จะแสดงเป็นแผนภาพโปรไฟล์ได้ ดังนี้คือ



แผนภาพที่ 3.2 เส้นโปรไฟล์ของประชากรกลุ่มที่ 1 และ 2 ที่ขนานกัน

ในแผนภาพที่ 2 เส้นโปรไฟล์ที่ขนานกันอาจจะทับกันสนิทหรือมีระดับเดียวกัน ทำการทดสอบในขั้นที่ 2

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

2. Test of coincident (Equal treatment levels) เป็นการทดสอบดูว่าถ้าหากเส้นโปรไฟล์มีการชนานกันจากการทดสอบในข้อที่ 1 แล้ว ระดับของทรีทเม้นท์จะเท่ากันหรือไม่ (เส้นโปรไฟล์จะทับกันสนิทหรือไม่) วิธีการทดสอบมีดังนี้

สมมติฐาน $H_o : J'\underset{\sim}{\mu}_1 = J'\underset{\sim}{\mu}_2$

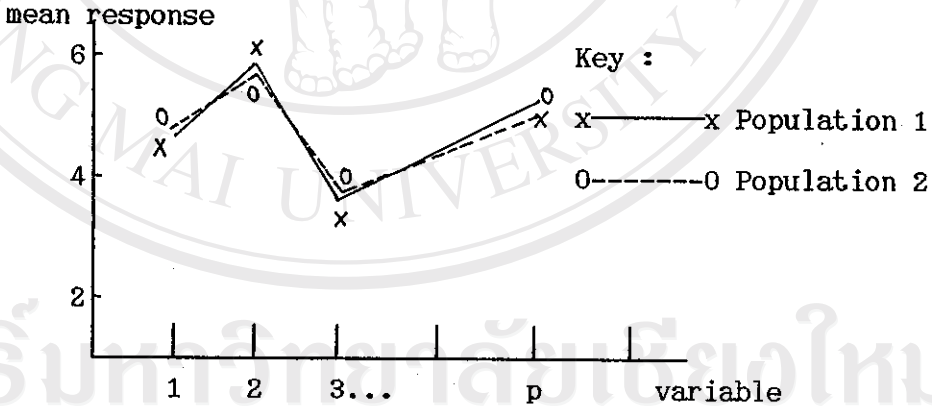
$J' = [1,\ldots,1]_{1\times p}$

ค่าสถิติที่ใช้ $$t_c = \frac{J'(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)}{\sqrt{J'SJ(\frac{1}{n_1} + \frac{1}{n_2})}}$$

จะยอมรับสมมติฐาน (นั่นคือ เส้นโปรไฟล์ทับกันสนิท) เมื่อ

$$|t_c| < t_{\alpha/2, n_1+n_2-2}$$

แสดงเป็นแผนภาพของเส้นโปรไฟล์ที่มีทรีทเม้นระดับเดียวกัน (ทับกันสนิท) ดังนี้ คือ



แผนภาพที่ 3.3 เส้นโปรไฟล์ของประชากรกลุ่มที่ 1 และ 2 ที่มีตัวแปรระดับเดียวกัน (ทับกันสนิท)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3. Test for equal response means เป็นการทดสอบดูว่าถ้าหากสมมติว่ามีการขนานกันของเส้นโปรไฟล์จากการทดสอบในข้อที่ 1 แล้ว ค่าเฉลี่ยของโปรไฟล์จะอยู่ในแนวระดับหรือไม่

สมมติฐาน $H_o : C(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2) = 0$

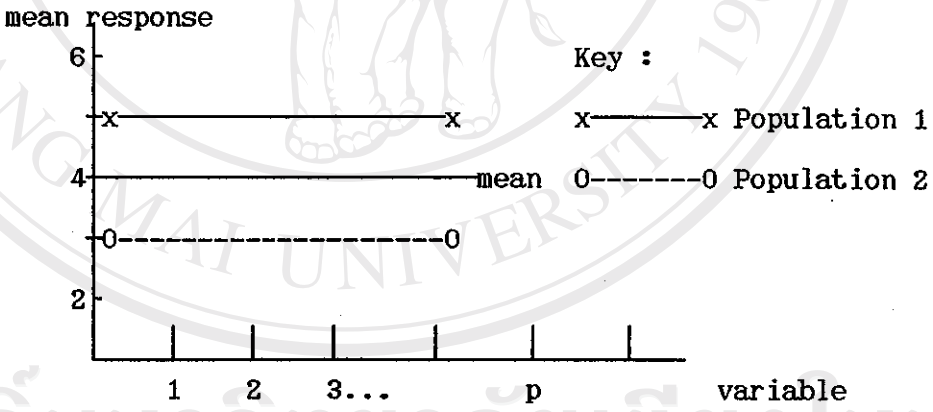
ค่าสถิติที่ใช้ $$T^2 = \frac{n_1 n_2}{n_1+n_2} [C(\bar{x}_1+\bar{x}_2)]' [CSC']^{-1} [C(\bar{x}_1+\bar{x}_2)]$$

$$F_c = \frac{(n_1+n_2-p)}{(n_1+n_2-2)(p-1)} \cdot T^2$$

จะยอมรับสมมติฐาน (นั่นคือ ค่าเฉลี่ยของโปรไฟล์อยู่ในแนวระดับ) เมื่อ

$$F_c < F_{\alpha ; (p-1),(n_1+n_2-p)}$$

แสดงเป็นแผนภาพของเส้นโปรไฟล์ที่มีค่าเฉลี่ยอยู่ในแนวระดับ ดังนี้ คือ



แผนภาพที่ 3.4 เส้นโปรไฟล์ของประชากรกลุ่มที่ 1 และ 2 ที่มีค่าเฉลี่ยอยู่ในแนวระดับ

หมายเหตุ โดยทั่วไปถ้าผลการวิเคราะห์พบว่า เส้นโปรไฟล์ทั้งสองขนานและทับกันสนิท รูปกราฟจะปรากฏเป็นเส้นตรงเส้นเดียว ซึ่งจะเป็นกรณีที่พบค่อนข้างยากและไม่บ่อยนัก

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.2 การวิเคราะห์จำแนกประเภท (Discriminant Analysis)

เทคนิคการวิเคราะห์จำแนกประเภทเป็นวิธีการทางสถิติที่อาศัยการเก็บรวบรวมข้อมูลซึ่งคาดว่าเกี่ยวข้องกับปัญหาที่จะจำแนกหน่วยวิเคราะห์ออกเป็นกลุ่ม เทคนิคดังกล่าวจะสามารถหาสมการที่ดีที่สุด โดยการลดความผิดพลาดในการจำแนกประเภทให้น้อยที่สุด และโดยการคัดตัวแปรที่สำคัญต่อการจำแนกกลุ่มมาใช้ในสมการนั้น ๆ จึงอาจกล่าวได้ว่าเทคนิคดังกล่าวนี้ นอกจากจะหาสมการที่ดีที่สุดจากข้อมูลที่มีอยู่ในการจำแนกหน่วยวิเคราะห์ไปตามกลุ่มต่าง ๆ แล้ว ยังสามารถกำหนดแยกตัวแปรที่จะจำแนกหน่วยวิเคราะห์ได้ดีที่สุดออกจากตัวแปรที่ช่วยการจำแนกกลุ่มได้น้อยกว่า

วัตถุประสงค์ของการวิเคราะห์จำแนกประเภท

วัตถุประสงค์สำคัญของการวิเคราะห์จำแนกประเภท คือ การสร้างสมการเพื่อแจกแจงหน่วยวิเคราะห์ว่าควรจัดเข้าอยู่ในกลุ่มใดจึงจะเหมาะสมและถูกต้อง

ลักษณะของข้อมูลที่ต้องการ

ข้อมูลที่ใช้ในการวิเคราะห์จำแนกประเภท ตัวแปรทุกตัวต้องมีการแจกแจงปกติ โดยอาจเป็นได้ทั้งมาตรนามบัญญัติหรือมาตรอันตรภาค ถ้าเป็นข้อมูลกลุ่มจะต้องแปรสภาพให้เป็นตัวแปรหุ่นก่อนที่จะทำการวิเคราะห์ สำหรับตัวแปรตามเป็นตัวแปรประเภทกลุ่มซึ่งทราบอยู่ก่อนแล้ว โดยอาจจะมี 2 กลุ่มหรือมากกว่าก็ได้

การสร้างสมการวิเคราะห์จำแนกประเภท

ถ้ามีตัวแปรอยู่ทั้งหมด p ตัวแปร และทุก ๆ หน่วยวิเคราะห์ในทุก ๆ กลุ่มก็ได้รับการเก็บรวบรวมข้อมูลด้วยตัวแปรทั้งหมดนี้

ให้

$X_1, X_2, X_3, \ldots, X_p$ เป็น p ตัวแปร

Y เป็น คะแนนจากการรวมกันของตัวแปรเหล่านั้น เราสามารถเขียนสมการจำแนกประเภทในรูปทั่วไปได้ดังนี้

$$Y = V_1X_1 + V_2X_2 + V_3X_3 + \ldots + V_pX_p$$

เมื่อ $V_1, V_2, V_3, \ldots, V_p$ เป็นน้ำหนักที่ไปคูณกับตัวแปรแต่ละตัว

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

การสร้างสมการวิเคราะห์จำแนกประเภทนี้สามารถทำได้หลายวิธี คือ วิธีตรง (direct method) วิธีแบบขั้นตอน (stepwise method) ฯลฯ

วิธีตรง

เป็นวิธีการที่ผู้วิจัยต้องการวิเคราะห์ตัวแปรทุกตัวที่ตนระบุ ด้วยเหตุผลทางทฤษฎีว่าจะแบ่งแยกได้กี่สมการ มีลักษณะอย่างไร โดยไม่ต้องการดูผลของแต่ละขั้นตอน เป็นวิธีการที่ประหยัดเวลา การใช้เครื่องคอมพิวเตอร์ และใช้หน่วยความจำน้อย

วิธีแบบขั้นตอน

เป็นวิธีการที่เลือกตัวแปรทีละตัวมาเข้าสมการ โดยหาตัวแปรที่ดีที่สุดในการจำแนกมาเข้าสมการเป็นตัวแรก จากนั้นก็จะหาตัวแปรที่ดีที่สุดตัวที่สองมาเข้าสมการเพื่อปรับปรุงแก้ไขทำให้สมการจำแนกดีขึ้น และในขั้นตอนต่อ ๆ ไป ก็จะเป็นการนำตัวแปรที่ดีที่สุดแต่ละตัวที่เหลือมาเข้าสมการ ในแต่ละขั้นตอน ตัวแปรที่ได้รับการคัดเลือกมาก่อนอาจถูกตัดทิ้งออกไปหากพบว่าเมื่อนำมารวมกับตัวแปรอื่น ๆ แล้ว ไม่ช่วยให้สมการจำแนกประเภทดีขึ้น จึงอาจกล่าวได้ว่า การคัดเลือกตัวแปรเข้ามาทีละขั้นตอนมีความคล้ายคลึงกับการวิเคราะห์ถดถอยพหุแบบขั้นตอน

การกำหนดค่าของโอกาสล่วงหน้า (prior probability)

การจำแนกกลุ่มขึ้นอยู่กับการกำหนดโอกาส (หรือความน่าจะเป็น) ของการเป็นสมาชิกกลุ่มใดกลุ่มหนึ่งที่ใช้ การกำหนดค่าของโอกาสล่วงหน้าสามารถทำได้ดังนี้

1. หากตัวอย่างของสมาชิกแต่ละกลุ่มมีจำนวนไม่เท่ากัน อาจใช้สัดส่วนของหน่วยวิเคราะห์ในแต่ละกลุ่มมาเป็นค่าของโอกาสที่หน่วยวิเคราะห์แต่ละหน่วยจะเป็นสมาชิกของกลุ่มนั้น ๆ

2. หากตัวอย่างของสมาชิกแต่ละกลุ่มมีจำนวนเท่า ๆ กัน แต่ในความเป็นจริงแล้วไม่เท่ากันอาจใช้สัดส่วนที่ทราบจากแหล่งอื่นมาเป็นค่าของโอกาสของความเป็นสมาชิกของแต่ละกลุ่มได้

3. เมื่อความเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งมีโอกาสเท่า ๆ กัน อาจให้ค่าของโอกาสความเป็นสมาชิกของแต่ละกลุ่มไว้เท่า ๆ กันได้

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

การทดสอบนัยสำคัญของสมการจำแนกประเภท

หลังจากการวิเคราะห์จำแนกประเภทอาจได้สมการจำแนกมาหลายสมการ ถ้าต้องการทราบว่าสมการใดมีอำนาจในการแบ่งกลุ่มได้อย่างมีนัยสำคัญทางสถิติ จะทำการทดสอบโดยมีสมมติฐาน ดังนี้

$H_o : \lambda_m > 0$

$H_1 : \lambda_m < 0$

และทดสอบได้โดยการใช้ค่า ไคสแควร์ จากสูตรต่อไปนี้

$V_m = [N-1-(P+K)/2] \ln(1+\lambda_m)$

เมื่อ N คือ จำนวนหน่วยวิเคราะห์ทั้งหมด

P คือ จำนวนตัวแปร

K คือ จำนวนกลุ่ม

โดยที่ V_m จะมีการแจกแจงเป็นแบบไคสแควร์ ที่มี degree of freedom = P+K-2m

($V_m \sim \chi^2_{\alpha, P+K-2m}$)

3.3 การหาค่าวินิจฉัย (diagnostic value) เป็นวิธีที่ใช้ประเมินผลเพื่อคัดแยกคนที่เป็นโรคและไม่เป็นโรคออกจากกัน โดยทำการประเมินผลดังนี้

3.3.1 sensitivity คือความสามารถในการทดสอบที่จะให้ค่าบวกที่แท้จริง (true positive) กี่เปอร์เซ็นต์ เมื่อนำค่าวินิจฉัยนี้ไปใช้ทดสอบกับพวกที่ทราบแน่นอนว่าเป็นโรคนั้น หรือนำไปตรวจสอบกับคนที่มีอาการเป็นโรคแล้วปรากฏว่าเป็นโรคนั้นจริง

Sensitivity : A good test gives positive results in diseased subjects (High true positives, few false negatives)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.3.2 specificity หมายถึงความจำเพาะของวิธีการที่ใช้ โดยจะให้ผลบวกเฉพาะกับสารที่ต้องการหาอย่างเดียวเท่านั้น การมีความจำเพาะต่ำทำให้เกิดผลบวกลวง (false positive) ในการทดสอบนั้น

หรือหมายถึงการนำค่าวินิจฉัยนี้ไปทดสอบแล้ว จะสามารถให้ค่าลบที่เป็นจริง (true negative) กี่เปอร์เซ็นต์ ถ้าไปทดสอบกับพวกที่ทราบแน่นอนว่าไม่เป็นโรคนั้น หรือในคนปกติ หรือนำไปตรวจสอบกับคนที่ไม่เป็นโรคแล้วได้ผลว่าไม่เป็นโรคนั้นจริง

Specificity : A good test gives negative results in healthy subjects (High true negatives, few false positives)

โดยทั่วไปแล้วค่าวินิจฉัยค่าหนึ่ง ควรจะได้ทำการทดสอบทั้ง Sensitivity และ Specificity ดังนี้คือ

(ก) ทำการทดสอบกับพวกที่ทราบแน่มาก่อนแล้วว่าเป็นโรคนั้นจริง เพื่อดูว่าค่าวินิจฉัยนั้นจะให้ผลบวกจริง (True Positive) และผลลบที่ผิด (False Negative) เท่าไร

(ข) ทำการทดสอบกับพวกที่ทราบแน่มาก่อนแล้วว่าไม่เป็นโรคนั้น เพื่อดูว่าค่าวินิจฉัยนั้นจะให้ผลลบจริง (True Negative) และผลบวกที่ผิด (False Positive) เท่าไร

เมื่อได้ค่าต่าง ๆ แล้ว จึงนำมาเสนอด้วยตาราง 4 ช่อง (Fourfold Table) เพื่อการคำนวนหาค่าต่าง ๆ ดังนี้

ค่าต่าง ๆ จากการทดสอบค่าวินิจฉัย

Test Mesurement or Result of Screening Test	True Mesuement or Stage of Disease		Total
	Positve	Negative	
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
Total	TP+FN	FP+TN	TP+FP+FN+TN

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

โดย

TP	คือ จำนวนคนเป็นโรคที่ค่าวินิจฉัยระบุถูกต้องว่าเป็นโรค
FN	คือ จำนวนคนเป็นโรคที่ค่าวินิจฉัยระบุผิดว่าไม่เป็นโรค
FP	คือ จำนวนคนไม่เป็นโรคที่ค่าวินิจฉัยระบุผิดว่าเป็นโรค
TN	คือ จำนวนคนไม่เป็นโรคที่ค่าวินิจฉัยระบุถูกต้องว่าไม่เป็นโรค
TP+FN	คือ จำนวนทั้งหมดของคนเป็นโรค
TN+FP	คือ จำนวนทั้งหมดของคนไม่เป็นโรค
TP+FN+FP+TN	คือ จำนวนคนทั้งหมดที่ได้รับการทดสอบ

การคำนวน ความสมบูรณ์ของค่าวินิจฉัย

$$1.\ \text{Sensitivity} = \frac{TP}{TP+FN} \times 100\ \%$$

$$2.\ \text{Specificity} = \frac{TN}{TN+FP} \times 100\ \%$$

ถ้าค่าวินิจฉัยให้ค่าทั้ง 2 สูงมาก ก็นับว่าเป็นค่าวินิจฉัยที่ดี แต่ไม่จำเป็นเสมอไปว่าค่าทั้ง 2 นี้จะได้ค่าสูงต่ำเหมือนกันหรือเท่ากัน

ค่าวินิจฉัยบางค่า อาจจะให้ค่าสูงทั้ง 2 ค่า

ค่าวินิจฉัยบางค่า อาจจะให้ค่าสูงเฉพาะ Sensitivity

ค่าวินิจฉัยบางค่า อาจจะให้ค่าสูงเฉพาะ Specificity

การที่นำค่าวินิจฉัยที่ให้ค่า Sensitivity และ Specificity สูงมาก ๆ ทั้งสองค่านั้น ไปทำการคัดแยกผู้เป็นโรคและไม่เป็นโรค ย่อมดีแน่ แต่บางครั้งก็อาจจะไม่ได้ค่าวินิจฉัยที่สมบูรณ์ตามต้องการได้ ดังนั้นในการเลือกใช้ค่าวินิจฉัยที่ให้ความสามารถในเรื่องใดมากหรือน้อยก็ต้องขึ้นอยู่กับผู้ใช้ว่า จะนำไปใช้ในเรื่องอะไร และหวังผลอย่างไร

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

<u>ตัวอย่าง</u> สมมติว่าเรามีเครื่องทดสอบ A ที่นำไปใช้ทดสอบผู้ป่วยที่เป็นโรคเบาหวาน (ซึ่งได้รับการวินิจฉัยด้วยการทดสอบหาน้ำตาลในโลหิตแล้ว) จำนวน 1,000 คน และในผู้ป่วยที่ไม่ได้เป็นโรคเบาหวานจำนวน 3,000 คน ได้ผลดังนี้ ให้ผลบวก 900 คน ในคนที่เป็นโรคเบาหวาน และ 1,200 คน ในคนที่ไม่เป็นโรคเบาหวาน

ดังนั้น จะได้ค่า

$$\text{Sensitivity} = \frac{900}{1000} \times 100 = 90\ \%$$

$$\text{Specificity} = \frac{1800}{3000} \times 100 = 60\ \%$$

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved