บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการวิจัยทางด้านการแพทย์ ด้านวิศวกรรมและอุตสาหกรรม ฯลฯ ส่วนใหญ่จะเป็นการศึกษาเกี่ยวกับเวลาที่เริ่มต้นศึกษาสิ่งที่สนใจไปจนกระทั่งเกิดเหตุการณ์ที่ต้องการศึกษาขึ้นกับสิ่งนั้น เรียกว่า เวลาการอยู่รอด (survival time, T)[1] เช่น เวลาที่เริ่มใช้เครื่องจักรใหม่ไปจนกระทั่งเครื่องจักรหยุดทำงาน เวลาที่เริ่มใช้หลอดไฟฟ้าใหม่จนกระทั่งหลอดเสีย เวลาที่คนไข้โรคมะเร็งในเม็ดเลือด เริ่มได้รับการรักษาจากโรงพยาบาล ไปจนกระทั่งเสียชีวิต ฯลฯ จากตัวอย่างคนไข้เป็นตัวอย่างการวิจัยทางด้านการแพทย์ เพื่อจะเปรียบเทียบผลของวิธีในการรักษาแบบใหม่กับแบบเดิม ดูว่าวิธีการรักษาแบบใหม่ ซึ่งคิดว่าดีกว่า จะให้เวลาการอยู่รอดนานกว่าแบบเดิมจริงหรือไม่ สิ่งที่สนใจคือ คนไข้โรคมะเร็งในเม็ดเลือด เวลาที่เริ่มต้นศึกษา คือ เวลาที่คนไข้เริ่มได้รับการรักษาจากโรงพยาบาล และเหตุการณ์ที่ต้องการศึกษา คือ การเสียชีวิตของคนไข้หลังจากได้รับการรักษา และในการศึกษาวิเคราะห์ผลก็ต้องพิจารณาปัจจัยหรือตัวแปรอิสระต่าง ๆ (independent variables, Z)[2] ที่เกี่ยวข้องกับคนไข้ เช่น อายุ จำนวนเม็ดเลือดขาวของคนไข้ ฯลฯ ดูว่ามีผลต่อเวลาการอยู่รอดหรือไม่ โดยปกติแล้ว เพื่อให้สะดวกและง่ายแก่การเข้าใจ จะให้เหตุการณ์ที่ต้องการศึกษาเป็นการสูญเสีย (failure) และเวลาที่เริ่มต้นศึกษาไปจนกระทั่งเกิดการสูญเสียเป็นเวลาการอยู่รอด (survival time) และการศึกษาวิเคราะห์ข้อมูลเวลาการอยู่รอดเพื่อตอบปัญหาต่าง ๆ ที่ต้องการทราบ ดังตัวอย่างข้างต้น เรียกว่า การวิเคราะห์การอยู่รอด (survival analysis)

---

[1] ตัวแปร T แทนเวลาการอยู่รอด

[2] ตัวแปร Z แทนตัวแปรอิสระต่าง ๆ ที่เกี่ยวข้อง

ดังนั้น จุดประสงค์โดยทั่วไปของการวิเคราะห์การอยู่รอดก็เพื่อตอบปัญหา ดังต่อไปนี้

1. เวลาการอยู่รอด (T) มีความแตกต่างกันระหว่างกลุ่มตัวอย่าง 2 กลุ่ม หรือมากกว่าหรือไม่ เช่น เวลาการอยู่รอดของกลุ่มคนไข้ที่ได้รับการรักษาแบบใหม่ ต่างจากกลุ่มคนไข้ที่ได้รับการรักษาแบบเดิมหรือไม่

2. เวลาการอยู่รอด ในกลุ่มที่ศึกษามีการแจกแจง (distribution) แบบใด

3. เวลาการอยู่รอด มีความสัมพันธ์กับตัวแปรอิสระ (Z) ต่าง ๆ หรือไม่ เช่น เพศ อายุ จำนวนเม็ดเลือดขาวของคนไข้ มีความสัมพันธ์หรือมีผลต่อเวลาการอยู่รอดของคนไข้หรือไม่

ซึ่งปัญหาต่าง ๆ เหล่านี้ สามารถตอบได้โดยใช้ฟังก์ชั่นของเวลาการอยู่รอด (function of survival time) อันได้แก่ ฟังก์ชั่นความหนาแน่น (density function), ฟังก์ชั่นการอยู่รอด (survival function) และฟังก์ชั่นการสูญเสีย (failure or hazard function) เช่น ถ้าทราบฟังก์ชั่นความหนาแน่นของเวลาการอยู่รอดก็สามารถบอกได้ว่า เวลาการอยู่รอดมีการแจกแจงแบบใด จากตัวอย่างข้างต้น สมมติฟังก์ชั่นความหนาแน่นของเวลาการอยู่รอดของคนไข้เป็น $f(t) = \lambda \exp(-\lambda t)$ แสดงว่า เวลาการอยู่รอดของคนไข้มีการแจกแจงแบบเอ็กโปเน็นเชียล ด้วยพารามิเตอร์ λ

ในทางปฏิบัติแล้ว ถ้าไม่ทราบค่าฟังก์ชั่นจริงก็ต้องใช้วิธีการประมาณค่าจากตัวอย่าง โดยปกติแล้ว ถ้าทราบการแจกแจงการอยู่รอด (survival distribution) ก็สามารถประมาณค่าฟังก์ชั่นของเวลาการอยู่รอดทั้ง 3 ฟังก์ชั่นข้างต้นได้ จากรูปแบบฟังก์ชั่นของการแจกแจง เป็นการประมาณค่าโดยวิธีพาราเมตริก (parametric) แต่ถ้าไม่ทราบการแจกแจงก็ต้องใช้การประมาณโดยวิธีนอนพาราเมตริก (nonparametric) แต่อย่างไรก็ตาม ทั้งวิธีพาราเมตริกและนอนพาราเมตริก นั้น โดยทั่วไปจะใช้กับข้อมูลที่มีค่าสังเกตทุกค่าเป็นค่าสังเกตสมบูรณ์ (complete observations) คือ ทราบเวลาเริ่มต้นศึกษาและทราบเวลาที่เกิดเหตุการณ์ที่

จะศึกษา หรือเวลาที่เกิดการสูญเสียของทุกหน่วยสังเกต เช่น ทราบเวลาที่คนไข้เริ่มได้รับการรักษาจากโรงพยาบาล และทราบเวลาที่คนไข้เสียชีวิต แต่ปัญหาก็คือ ในทางปฏิบัติมีข้อมูลเป็นจำนวนมากที่มีค่าสังเกตบางค่าเป็นค่าสังเกตไม่สมบูรณ์ (incomplete or censored observations) คือ ทราบเวลาเริ่มต้นศึกษา แต่ไม่ทราบเวลาที่เกิดการสูญเสียของบางหน่วยสังเกต ซึ่งอาจเกิดจากไม่สามารถติดตามหน่วยสังเกตได้ หรือ หน่วยสังเกตยังไม่เกิดการสูญเสีย เมื่อสิ้นสุดการศึกษา เพื่อที่จะเริ่มทำการวิเคราะห์ ฯลฯ เช่น คนไข้ขาดการติดต่อกับโรงพยาบาลหลังจากที่ได้รับการรักษาแล้ว หรือคนไข้ยังไม่เสียชีวิตเมื่อสิ้นสุดการศึกษา เป็นต้น และข้อมูลส่วนใหญ่จะไม่ทราบการแจกแจงด้วย ดังนั้นในการประมาณค่าฟังก์ชั่นการอยู่รอด สำหรับข้อมูลชนิดนี้ จำเป็นต้องใช้การประมาณโดยวิธีนอนพาราเมตริก ที่เป็นวิธีเฉพาะที่เหมาะสมกับชนิดของข้อมูล ซึ่งในปัจจุบันก็ได้มีการพัฒนาวิธีการขึ้นมาใช้หลายวิธี วิธีที่สำคัญและนำมาใช้กันมาก ได้แก่

1. วิธี Product-Limit (PL) เป็นวิธีที่พัฒนาขึ้น โดย Kaplan & Meier (1958)
2. วิธี Life-table (actuarial) โดย Berkson & Gage (1950) เป็นวิธีทางสถิติที่เก่าแก่และนำมาใช้กันอย่างกว้างขวาง
3. วิธี Cox's regression model ของ Cox (1972)

วิธีการประมาณค่าฟังก์ชั่นการอยู่รอด 3 วิธีข้างต้นนี้ เป็นวิธีนอนพาราเมตริก สำหรับข้อมูลที่มีค่าสังกตไม่สมบูรณ์โดยเฉพาะ และในการศึกษาครั้งนี้ ผู้วิจัยได้นำวิธีการประมาณค่าฟังก์ชั่นทั้ง 3 วิธี มาศึกษากับปัญหา 2 ตัวอย่าง (two-sample problem)[3] คือ ศึกษาปัญหาที่ข้อมูลแบ่งเป็น 2 กลุ่มตัวอย่างย่อย หรือได้รับกรรมวิธี 2 กรรมวิธี เช่น ข้อมูลคนไข้โรค

---

[3] D.R. Cox, "Regression model and Life-tables". Journal of the Royal Statistical Society, (Series B34, 1972), p. 193.

มะเร็งในเม็ดเลือด แบ่งเป็น 2 กลุ่มย่อย โดยที่กลุ่มหนึ่งได้รับการรักษาแบบใหม่ ส่วนอีกกลุ่มได้รับการรักษาแบบเดิม และข้อมูลที่ศึกษาเป็นข้อมูลที่มีค่าสังเกตไม่สมบูรณ์ที่เป็นแบบผสมของ Singly failure censored และ multiply time censored[4] คือ ข้อมูลที่มีจำนวนค่าสังเกตไม่สมบูรณ์เท่ากับจำนวนที่กำหนดขึ้น และค่าสังเกตไม่สมบูรณ์แต่ละค่าอาจจะมีค่าเท่ากับค่าสังเกตสมบูรณ์ ที่มีค่ามากที่สุด ซึ่งเกิดจากหน่วยสังเกตยังไม่เกิดการสูญเสียเมื่อสิ้นสุดการศึกษา หรืออาจจะมีค่าไม่เท่ากับค่าสังเกตใด ๆ เลยก็ได้ ค่าไม่สมบูรณ์นี้อาจจะเกิดจากหน่วยสังเกตขาดการติดต่อหรืออาจเกิดจากสาเหตุอื่น ๆ

ในการนำวิธีการประมาณค่าฟังก์ชั่นทั้ง 3 วิธี มาใช้กับข้อมูลที่เป็นปัญหา 2 ตัวอย่าง และมีค่าสังเกตไม่สมบูรณ์ดังกล่าวข้างต้น ก็เพื่อศึกษาดูว่าในข้อมูลตัวอย่างชุดเดียวกัน ค่าประมาณของฟังก์ชั่นการอยู่รอดที่ประมาณ โดยวิธีทั้ง 3 วิธี (PL, Life-table, Cox's regression) จะมีค่าแตกต่างกันหรือไม่ วิธีใดจะให้ค่าประมาณต่ำสุดและสูงสุด และถ้าข้อมูลมีจำนวนค่าสังเกตไม่สมบูรณ์เพิ่มมากขึ้น ค่าประมาณของฟังก์ชั่นจะเปลี่ยนแปลงหรือไม่ วิธีที่ให้ค่าต่ำสุดและสูงสุด อาจจะเปลี่ยนไป โดยจะพิจารณาเฉพาะค่ามัธยฐานการอยู่รอด (median survival time หรือ เปอร์เซนต์ไทล์ที่ 50) ซึ่งประมาณจากฟังก์ชั่นการอยู่รอดของแต่ละวิธี ทั้งนี้เพราะว่า การแจกแจงการอยู่รอดโดยทั่วไปแล้ว มักจะเบ้ (skewed) การใช้ค่ามัธยฐานจึงดีกว่าค่าเฉลี่ย (mean)[5] ดังนั้น ถ้าค่าประมาณมัธยฐานของวิธีใดมีค่าใกล้เคียงกับค่ามัธยฐานจริง ก็ควรจะเลือกใช้วิธีการประมาณนั้น และทำการทดสอบการเปรียบเทียบการแจกแจงการอยู่รอด (survival distribution) ของ 2 กลุ่มตัวอย่างย่อย เพื่อดูว่าผลการทดสอบจะเปลี่ยนแปลงหรือไม่ ถ้าจำนวนค่าสังเกตไม่สมบูรณ์ในตัวอย่างเพิ่มมากขึ้น

---

[4] W.Nelson, Applied Life Data Analysis, (John Wiley & Sons: New York, 1982), pp. 7-9.

[5] Elisa T. Lee, Statistical Methods for Survival Data Analysis. (Lifetime Learning Publication : USA, 1980), p. 11.

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีนอนพาราเมตริก สำหรับการประมาณค่าฟังก์ชันการอยู่รอดในปัญหา 2 ตัวอย่าง ที่มีค่าสังเกตไม่สมบูรณ์ (two-sample problem with censored observations) โดยวิธี Product-Limit (PL). วิธี Life-table (actuarial) และวิธี Cox's regression model

2. เพื่อศึกษาเปรียบเทียบวิธีการประมาณค่าฟังก์ชันการอยู่รอดในข้อ 1

3. เพื่อศึกษาดูว่า เมื่อมีจำนวนค่าสังเกตไม่สมบูรณ์ในตัวอย่างเพิ่มมากขึ้น ค่าประมาณของแต่ละวิธีในข้อ 1 และผลการทดสอบการเปรียบเทียบการแจกแจงการอยู่รอดของ 2 กลุ่มตัวอย่างย่อย จะเปลี่ยนแปลงหรือไม่

1.3 สมมติฐานในการวิจัย

สมมติฐานในการวิจัยครั้งนี้ คือ วิธี Cox's regression model เป็นวิธีการประมาณค่าฟังก์ชันการอยู่รอดที่ดีกว่า วิธี Product-Limit และวิธี Life-table

1.4 ข้อตกลงเบื้องต้น

1. ตัวแปร T (เวลาการอยู่รอด) มีค่าต่อเนื่อง (continuous)

2. ข้อมูลไม่มีค่าสังเกตซ้ำ (tied observations)

3. ตัวแปรอิสระ Z เป็นตัวแปรคงที่ (fixed variables) ซึ่งเป็นตัวแปรเชิงปริมาณ (quantitative variables) หรือถ้าเป็นตัวแปรเชิงคุณภาพ (qualitative variables) ก็ต้องแปลงให้เป็นตัวแปรหุ่น (dummy variables) ก่อนการนำไปทำการวิเคราะห์[6]

---

[6] Ibid., pp. 299-300.

1.5 ขอบเขตของการวิจัย

1. ศึกษาวิธีการและทฤษฎีที่ใช้ในการประมาณค่าฟังก์ชั่นการอยู่รอดโดยวิธี Product -Limit (PL), วิธี Life-table และวิธี Cox's regression model พร้อมทั้งศึกษาทฤษฎีที่พัฒนาไปใช้สำหรับกรณีเฉพาะในปัญหา 2 ตัวอย่างที่มีค่าสังเกตไม่สมบูรณ์

2. ศึกษาเปรียบเทียบค่าประมาณฟังก์ชั่นการอยู่รอดของแต่ละวิธีในข้อ 1 ในปัญหา 2 ตัวอย่าง ที่มีค่าสังเกตไม่สมบูรณ์ที่เป็นแบบผสมของ singly failure และ multiply time censored เท่านั้น

3. ตัวแปรอิสระ Z มีเพียงตัวเดียว เพราะเป็นการศึกษาในปัญหา 2 ตัวอย่าง โดยไม่มีตัวแปรอิสระอื่นเข้ามาเกี่ยวข้อง เป็นตัวแปรหุ่น มีค่าดังนี้

$$Z_{1i} = \begin{cases} 0 & \text{เมื่อค่าสังเกตที่ } i \text{ อยู่ในกลุ่มตัวอย่างย่อยที่ 1} \\ 1 & \text{เมื่อค่าสังเกตที่ } i \text{ อยู่ในกลุ่มตัวอย่างย่อยที่ 2} \end{cases}$$

4. ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่สร้างขึ้น (random number generation) ให้มีการแจกแจงแบบเอ็กโปเน็นเชียล (exponential distribution) ด้วยพารามิเตอร์ $\lambda = 0.10$ ($\mu = 10$, $\sigma^2 = 100$)

จาก[7] y_p (100pth percentile) $= -\mu \ln(1-p)$

ดังนั้น $y_{.5}$ (50 th percentile) $= -10 \ln(0.5) \sim 6.93$

5. ข้อมูลตัวอย่างที่ใช้ในการศึกษาจะสร้างขึ้นกลุ่มตัวอย่างละ 100 ครั้ง

6. ขนาดตัวอย่างที่ใช้ในการศึกษา n = 10, 30, 50 แต่ละกลุ่มตัวอย่างแบ่งเป็น 2 กลุ่มตัวอย่างย่อย มีขนาดเท่ากัน และมีจำนวนค่าสังเกตไม่สมบูรณ์เท่ากับ 10, 20 และ 30 เปอร์เซนต์ของขนาดตัวอย่าง n โดยการสุ่ม

7. การศึกษาการเปรียบเทียบการแจกแจงการอยู่รอดของ 2 กลุ่มตัวอย่างย่อย จะใช้ Cox-Mantel test $\alpha = 0.05, 0.01$

---

[7] Nelson, Applied Life Data Analysis. p.22.

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้ทราบถึงวิธีนอนพาราเมตริก สำหรับการประมาณค่าฟังก์ชั่นการอยู่รอดที่เหมาะสมกับข้อมูลที่มีค่าสังเกตไม่สมบูรณ์ โดยเฉพาะกรณีปัญหา 2 ตัวอย่าง

2. สามารถนำวิธีที่เหมาะสมไปใช้กับข้อมูลจริง ที่มีค่าสังเกตไม่สมบูรณ์ในกรณีปัญหา 2 ตัวอย่าง และนำไปพัฒนาใช้กับกรณีหลายตัวอย่างได้