บทที่ 2

สถิติที่ใช้ในการวิจัย

โดยทั่วไปในการวิเคราะห์ข้อมูลเกี่ยวกับระยะเวลาที่ยังคงอยู่ (Lifetime) มักจะพบกับปัญหาว่าข้อมูลที่เก็บรวบรวมมาได้เป็นข้อมูลที่ไม่สมบูรณ์ ทั้งนี้เนื่องจากข้อจำกัดต่างๆ ที่เกิดขึ้นในการเก็บรวบรวมข้อมูล โดยจะพบว่าข้อมูลที่ไม่สมบูรณ์จะมีค่าถูกตัดทิ้งทางขวา ทั้งนี้ในการพิจารณาการเกิดข้อมูลที่มีค่าถูกตัดทิ้งนั้นสามารถแยกชนิดได้ตามวิธีการได้มาของข้อมูลซึ่งแบ่งได้เป็น 3 ชนิด คือ ถ้าข้อมูลได้จากการทดลองภายใต้การกำหนดเวลาที่จะสิ้นสุดการเก็บข้อมูลของแต่ละหน่วยตัวอย่างไว้ล่วงหน้า โดยที่เวลาที่ได้กำหนดไว้นี้เป็นเวลาเดียวกันแล้ว ข้อมูลที่ได้นี้จะเป็นข้อมูลที่ถูกตัดทิ้งประเภทที่ 1 (Type I Censoring) นั่นคือ จะสามารถทราบข้อมูลระยะเวลาที่ยังคงอยู่ที่แท้จริงได้เฉพาะกรณีที่ข้อมูลมีค่าน้อยกว่าเวลาที่ถูกกำหนดไว้เท่านั้น ซึ่งพบว่าจำนวนข้อมูลที่ทราบค่าแท้จริงนี้จะไม่แน่นอน แต่ถ้ามีการกำหนดจำนวนดังกล่าวนี้ไว้ล่วงหน้าอย่างแน่นอนแล้วข้อมูลที่ได้จะเป็นข้อมูลที่ถูกตัดทิ้งประเภทที่ 2 (Type II Censoring) โดยข้อมูลประเภทนี้จะไม่ทราบเวลาที่ใช้ในการเก็บรวบรวมข้อมูลที่แน่นอน เพราะจะต้องดำเนินการทดลองต่อไปจนกระทั่งสามารถบันทึกข้อมูลได้ครบตามจำนวนที่ได้กำหนดไว้ล่วงหน้า ส่วนข้อมูลที่ถูกตัดทิ้งชนิดที่ 3 คือ ข้อมูลที่ถูกตัดทิ้งแบบสุ่ม (Random Censoring) ซึ่งเป็นลักษณะของข้อมูลที่มีการตัดข้อมูลในระหว่างการทดลองและภายหลังสิ้นสุดการทดลอง

สำหรับในงานวิจัยครั้งนี้ ผู้วิจัยได้ทำการศึกษากรณีที่ข้อมูลถูกตัดทิ้งประเภทที่ 1 เท่านั้น โดยศึกษาภายใต้การแจกแจงแบบไวบูลล์ และการแจกแจงแบบลอกนอร์มอล โดยการนำข้อมูลมาทำการวิเคราะห์ทางสถิติ เพื่อประมาณค่าฟังก์ชันการอยู่รอดตามวิธีต่างๆ ดังนั้นในบทนี้จะกล่าวถึงรายละเอียดเกี่ยวกับการแจกแจงการอยู่รอด และฟังก์ชันต่างๆ รวมทั้งวิธีการประมาณฟังก์ชันการอยู่รอดโดยวิธีการประมาณที่ไม่ใช้พารามิเตอร์ ซึ่งมีรายละเอียดดังนี้

<u>การแจกแจงการอยู่รอด</u>

ให้ T เป็นตัวแปรสุ่มเวลาที่จะเกิดเหตุการณ์ที่สนใจศึกษา หรือเวลาการอยู่รอด

$f(t)$ เป็นฟังก์ชันความหนาแน่น (Density Function)

$F(t)$ เป็นฟังก์ชันการแจกแจง (Distribution Function)

โดยที่ $F(t) = \Pr(T \le t)$

$S(t)$ เป็นฟังก์ชันการอยู่รอด (Survival Function) มีค่าเท่ากับความน่าจะเป็นที่หน่วยตัวอย่างมีการอยู่รอดนานกว่าเวลา t นั่นคือ

$$S(t) = \Pr(\text{หน่วยตัวอย่างมีการอยู่รอดนานกว่าเวลา } t)$$
$$= \Pr(T > t)$$
$$= 1 - F(t)$$

โดยที่ $S(t)$ มีคุณสมบัติดังนี้

1. $S(t)$ เป็นฟังก์ชันไม่เพิ่ม (Nonincreasing function)
2. $S(t)$ เป็นฟังก์ชันต่อเนื่องของ t
3. $S(t) = 1$ เมื่อ $t = 0$
4. $S(t) = 0$ เมื่อ $t = \infty$

$h(t)$ เป็นฟังก์ชันภาวะภัย (Hazard Function) ซึ่งมีค่าเท่ากับลิมิตของความน่าจะเป็นที่หน่วยตัวอย่างจะสูญเสียในช่วงเวลาสั้นๆ จาก t ถึง $t + \Delta t$ ต่อหน่วยเวลา Δt เมื่อแต่ละหน่วยตัวอย่างมีการอยู่รอดถึงเวลา t นั่นคือ

$$h(t) = \lim_{\Delta t \to 0} \Pr\{\text{หน่วยตัวอย่างที่มีอายุ t จะสูญเสียในช่วง } (t, t + \Delta t)\} / \Delta t$$

โดยที่ $h(t)$ มีคุณสมบัติดังนี้

1. $h(t) \ge 0 \quad ; -\infty < t < \infty$
2. $\lim_{t \to -\infty} \int_{-\infty}^{t} h(t)dt = 0$ และ $\lim_{t \to \infty} \int_{-\infty}^{t} h(t)dt = \infty$

$E(T)$ เป็นค่าเฉลี่ยของตัวแปรสุ่ม T

สามารถเขียนความสัมพันธ์ระหว่างฟังก์ชันการอยู่รอด ฟังก์ชันความหนาแน่นฟังก์ชันการแจกแจง และฟังก์ชันภาวะภัยได้ดังนี้

1. $f(t) = F'(t) = -S'(t)$
2. $h(t) = \dfrac{f(t)}{S(t)} = -\dfrac{S'(t)}{S(t)}$
3. $H(t) = \int_0^t h(u)du = -\ln S(t)$

หรือ $S(t) = \exp(-\int_0^t h(u)du) = \exp(-H(t))$

สำหรับในงานวิจัยนี้พิจารณารูปแบบการแจกแจงการอยู่รอด 2 รูปแบบ คือ

1. <u>การแจกแจงแบบไวบูลล์</u> (Weibull Distribution)

เมื่อตัวแปรสุ่ม T มีการแจกแจงแบบไวบูลล์แล้ว สามารถเขียนฟังก์ชันต่างๆ ได้ดังนี้

$$f(t) = \begin{cases} \lambda\gamma(\lambda t)^{\gamma-1}\exp\left(-(\lambda t)^{\gamma}\right) & 0 < t < \infty, \lambda > 0, \gamma > 0 \\ 0 & \text{อื่นๆ} \end{cases}$$

$$F(\mathrm{t}) = 1 - \exp\left(-(\lambda \mathrm{t})^{\gamma}\right)$$

$$S(\mathrm{t}) = \exp\left(-(\lambda \mathrm{t})^{\gamma}\right)$$

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$$

$$E(T) = \frac{\Gamma(1+1/\gamma)}{\lambda}$$

เมื่อ $\Gamma(\nu)$ เป็นฟังก์ชันแกมมา (Gamma Function) ซึ่งสามารถคำนวณได้ดังนี้

$$\begin{aligned} \Gamma(\nu) &= \int_0^{\infty} y^{\nu-1}e^{-y}dy \\ &= (\nu-1)\Gamma(\nu-1) \\ &= (\nu\text{-}1)! \qquad \text{เมื่อ } \nu \text{ เป็นจำนวนเต็ม} \end{aligned}$$

และ $$\Gamma(.5) = \sqrt{\pi}$$

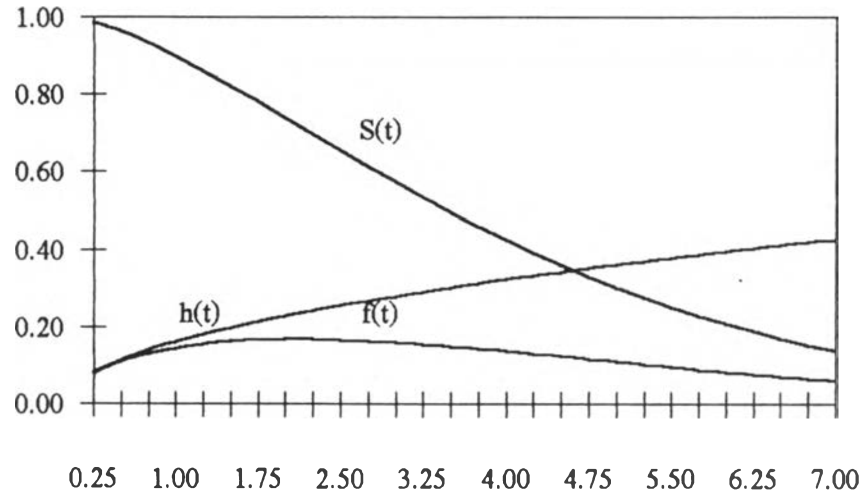
โดยที่ เมื่อกำหนดให้พารามิเตอร์ λ คงที่แล้ว ลักษณะรูปแบบของฟังก์ชันภาวะภัยจะขึ้นอยู่กับพารามิเตอร์ γ ดังนี้

ถ้า $\gamma <$ 1.0 แล้ว ฟังก์ชันภาวะภัย $h(t)$ จะลดลงเมื่อ t เพิ่มขึ้น

$\gamma =$ 1.0 แล้ว ฟังก์ชันภาวะภัย $h(t)$ จะคงที่เมื่อ t เพิ่มขึ้น

$\gamma >$ 1.0 แล้ว ฟังก์ชันภาวะภัย $h(t)$ จะเพิ่มขึ้นเมื่อ t เพิ่มขึ้น

สำหรับรูปแบบของฟังก์ชัน $f(t), S(t)$และ $h(t)$ เมื่อการแจกแจงการอยู่รอดเป็นแบบไวบูลล์ที่มีค่าเฉลี่ยของการแจกแจง $E(t)$ เป็น 4.0 และ $\gamma = 1.5$ แสดงได้ดังรูปที่ 2.1



รูปที่ 2.1

2. การแจกแจงแบบลอกนอร์มอล (Lognormal Distribution)

เมื่อตัวแปรสุ่ม T มีการแจกแจงแบบลอกนอร์มอลแล้ว สามารถเขียนฟังก์ชันต่างๆ ได้ดังนี้

$$f(t) = \begin{cases} \dfrac{1}{t\sigma\sqrt{2\pi}}\exp\left[-\dfrac{(\ln(t)-\mu)^2}{2\sigma^2}\right] & t>0, -\infty<\mu<\infty, \sigma>0 \\ 0 & \text{อื่นๆ} \end{cases}$$

$$F(t) = \Phi\left[\frac{\ln(t)-\mu}{\sigma}\right]$$

เมื่อ $\Phi(x)$ คือ ฟังก์ชันการแจกแจงปกติมาตรฐาน

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x \exp\left(-\frac{u^2}{2}\right)du$$

ดังนั้น

$$S(t) = 1-\Phi\left[\frac{\ln(t)-\mu}{\sigma}\right]$$

$$E(T) = \exp\left(\mu+\frac{\sigma^2}{2}\right)$$

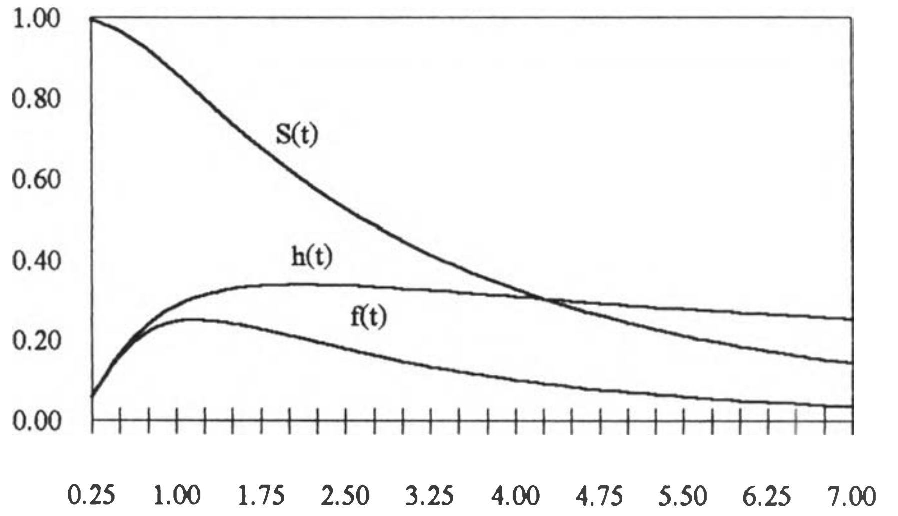
เมื่อกำหนดให้ μ คงที่แล้ว ลักษณะรูปแบบของฟังก์ชันภาวะภัยจะขึ้นอยู่กับ σ ดังนี้

ถ้า $\sigma \geq 2.0$ แล้วฟังก์ชันภาวะภัย $h(t)$ จะลดลงเมื่อ t เพิ่มขึ้น

$\sigma < 0.5$ แล้วฟังก์ชันภาวะภัย $h(t)$ จะเพิ่มขึ้นเมื่อ t เพิ่มขึ้น

σ ใกล้เคียงกับ 1.0 แล้วฟังก์ชันภาวะภัย $h(t)$ ค่อนข้างจะคงที่ โดยจะเพิ่มขึ้นอย่างรวดเร็วในช่วงต้นๆ และจะคงที่เมื่อ t เพิ่มขึ้น

สำหรับรูปแบบของฟังก์ชัน $f(t), S(t)$ และ $h(t)$ เมื่อการแจกแจงการอยู่รอดเป็นแบบลอกนอร์มอลที่มีค่าเฉลี่ยของการแจกแจง $E(t)$ เป็น 4.0 และ $\sigma = 0.9$ แสดงได้ดังรูปที่ 2.2



รูปที่ 2.2

<u>สถิติที่ใช้ในการวิจัย</u>

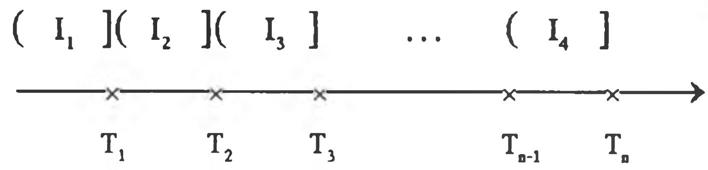
1. วิธีพีแอล (Product Limit (PL) Method)[1]

วิธีพีแอล เป็นวิธีการประมาณฟังก์ชันการอยู่รอดสำหรับข้อมูลที่ไม่สมบูรณ์ ซึ่งถูกพัฒนาขึ้นโดย แค็พแลนและไมเออร์ (Kaplan and Meier, 1958) ซึ่งมีแนวความคิดและรายละเอียดในการคำนวณ ดังนี้

สมมติว่า ค่าสังเกตของเวลาที่อยู่รอด (Survival Time) จำนวน n ตัวอย่าง มีค่าสังเกตเป็น $X_1, X_2, \ldots, X_n$ นำค่าสังเกตมาเรียงลำดับค่าจากน้อยไปมากจะได้ $T_1 < T_2 < \ldots < T_n$

---

[1] Kaplan EL, and Meier P. "Nonparametric Estimation From Incomplete Observations," <u>Journal of the American Statistical Association</u> 53 (1958) : 457-481.



ให้ n_i เป็นจำนวนค่าสังเกตที่อยู่รอด ณ จุดเริ่มต้นของช่วงเวลา I_i

d_i เป็นจำนวนตัวอย่างที่เสียชีวิตระหว่างช่วงเวลาที่ I_i

p_i เป็นความน่าจะเป็นที่จะอยู่รอดถึงช่วงเวลา I_i โดยที่อยู่รอดมาแล้ว ณ จุดเริ่มต้นของช่วงเวลา I_i นั่นคือ

$$p_i = \Pr(T>t_i / T>t_{i-1})$$

$$q_i = 1 - p_i$$

โดยที่ $$\hat{q}_i = \frac{d_i}{n_i}$$

เมื่อความน่าจะเป็นของการมีชีวิตอยู่รอด t_k ปี คือ

$$\begin{aligned} S(t_k) &= \Pr(T > t_k) \\ &= \Pr(T > t_1) * \Pr(T > t_2 | T > t_1) * \cdots * \Pr(T > t_k | T > t_{k-1}) \\ &= p_i \cdot p_2 \cdots p_k \end{aligned}$$

ดังนั้น สำหรับกรณีทั่วไป

$$\begin{aligned} \hat{S}(t) &= \prod_{\{i;\, t_i \le t\}} \hat{p}_i \\ &= \prod_{\{i;\, t_i \le t\}} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} \end{aligned} \qquad (2.1)$$

และถ้าข้อมูลไม่มีค่าสังเกตซ้ำ (Tied Observations) ในช่วงเวลา I_i

จะได้ว่า $$\hat{q}_i = \frac{1}{n_i}$$

$$\hat{p}_i = 1 - \hat{q}_i$$

$$\hat{p}_i = 1 - \frac{1}{n_i}$$ เมื่อ $\delta_i = 1$ ข้อมูลไม่ถูกตัดทิ้ง

$$= 1$$ เมื่อ $\delta_i = 0$ ข้อมูลถูกตัดทิ้ง

ดังนั้น
$$\hat{S}(t) = \prod_{\{i;\, t_i \le t\}} \hat{p}_i$$
$$= \prod_{\{i;\, t_i \le t\}} \left(1 - \frac{1}{n_i}\right)^{\delta_i}$$
$$= \prod_{\{i;\, t_i \le t\}} \left(1 - \frac{1}{n-i+1}\right)^{\delta_i}$$
$$\hat{S}_{PL}(t) = \prod_{\{i;\, t_i \le t\}} \left(\frac{n-i}{n-i+1}\right)^{\delta_i} \qquad (2.2)$$

เมื่อ n เป็นจำนวนข้อมูลทั้งหมดที่ถูกตัดทิ้งและไม่ถูกตัดทิ้ง

i เป็นลำดับที่ของข้อมูลในการเรียงลำดับค่าจากน้อยไปมาก

2. วิธีฟังก์ชันภาวะภัย (Hazard Function Method)[2]

วิธีนี้จะประมาณฟังก์ชันการอยู่รอดจากการประมาณฟังก์ชันภาวะภัยสะสม (Cumulative Hazard Function; $\Lambda(t)$) ก่อน โดยอาศัยความสัมพันธ์ระหว่างฟังก์ชันการอยู่รอดและฟังก์ชันภาวะภัยสะสม ดังนี้

เนื่องจาก
$$h(t) = \frac{f(t)}{1-F(t)} = \frac{F'(t)}{1-F(t)} = -\frac{S'(t)}{S(t)}$$
$$\Lambda(t) = \int_0^t h(u)du$$
$$= \int_0^t -\frac{S'(t)}{S(t)}du$$
$$= -\ln S(t)$$

ดังนั้น
$$S(t) = e^{-\Lambda(t)} \qquad (2.3)$$

และ
$$\hat{\Lambda}(t) = -\ln \hat{S}(t) \qquad (2.4)$$

และจากการประมาณ $\hat{S}(t)$ ด้วยวิธีพีแอลในสมการที่ (2.1) และสมการ (2.4) จะได้

$$\hat{\Lambda}(t) = -\ln\left[\prod_{i:\, t_i \le t}\left(1 - \frac{d_i}{n_i}\right)^{\delta_i}\right] \qquad (2.5)$$

---

[2] Rupert G. Miller, Jr., Survival Analysis New York : John Wiley & sons,1981.

$$\hat{\Lambda}(t) \quad = \quad -\delta_i \sum_{i:t_i \le t} \ln\left(1 - \frac{d_i}{n_i}\right) \tag{2.6}$$

และเพราะว่า $\quad -\ln\left(1-\frac{d_i}{n_i}\right) \quad = \quad \frac{d_i}{n_i} + \frac{1}{2}\left(\frac{d_i}{n_i}\right)^2 + \cdots$

เมื่อพิจารณาเฉพาะพจน์ที่มีกำลังเป็น 1 ดังนั้นจะได้ว่า

$$-\ln\left(1-\frac{d_i}{n_i}\right) \quad \approx \quad \frac{d_i}{n_i}$$

จาก (2.6) จะสามารถประมาณฟังก์ชันภาวะภัยสะสมได้เป็น

$$\hat{\Lambda}(t) \quad = \quad -\delta_i \sum_{i:t_i \le t} \frac{d_i}{n_i} \tag{2.7}$$

ดังนั้น จากสมการ (2.3) และ (2.7) จะได้ตัวประมาณฟังก์ชันการอยู่รอด ซึ่งเรียกว่าตัวประมาณแนลสัน-อาเลน (Nelson-Aalen Estimator) เป็น

$$\hat{S}(t) \quad = \quad \exp\left(-\delta_i \sum_{i:t_i \le t} \frac{d_i}{n_i}\right) \tag{2.8}$$

และเมื่อข้อมูลไม่มีค่าสังเกตซ้ำ จะได้

$$\hat{\Lambda}(t) \quad = \quad \sum_{i:t_i \le t} \frac{\delta_i}{n-i+1} \tag{2.9}$$

ดังนั้น จากสมการที่ (2.3) และ (2.9) จะได้ตัวประมาณฟังก์ชันการอยู่รอด คือ

$$\hat{S}_{HA}(t) \quad = \quad \exp\left(-\sum_{i:t_i \le t} \frac{\delta_i}{n-i+1}\right) \tag{2.10}$$

3. วิธีเบส์ (Bayes Method)[3]

วิธีเบส์นี้ เป็นวิธีทางสถิติที่ต้องอาศัยข้อมูลจากกลุ่มตัวอย่างที่นำมาศึกษา และข้อมูลหรือประสบการณ์ในอดีตเข้าช่วยในการประมาณค่า โดยจากข้อมูลหรือประสบการณ์ในอดีตนี้ทำให้สามารถประมาณการแจกแจงก่อน (Prior Distribution) ได้ และเมื่อทราบการแจกแจงก่อน และข้อมูลจากกลุ่มตัวอย่างที่นำมาศึกษาแล้ว จะสามารถทราบการแจกแจงภายหลัง (Posterior Distribution) ได้ โดยจะนำการแจกแจงดังกล่าวนี้ไปใช้ในการประมาณค่าที่ต้องการได้

สำหรับการวิจัยครั้งนี้ กำหนดให้ $T_1, T_2, \ldots, T_n$ เป็นตัวอย่างสุ่มจากฟังก์ชันการแจกแจงสะสม (Cumulative Distribution Function) F และ Tc เป็นค่าคงที่ซึ่งทราบค่า และให้ $Z_i = \min(T_i, Tc)$ เมื่อ $i = 1, 2, \ldots, n$

ในที่นี้จะหาตัวประมาณเบส์ภายใต้การกำหนดให้ฟังก์ชันความสูญเสียเป็นความคลาดเคลื่อนกำลังสอง (squared-error loss function)

$$l\left(F(t)-\hat{F}(t)\right) = \left(F(t)-\hat{F}(t)\right)^2$$

ดังนั้น ตัวประมาณเบส์สามารถหาได้จากการหาค่าคาดหวังของการแจกแจงภายหลังของ F เมื่อกำหนด (Z_i, δ_i)

สำหรับการประมาณค่าฟังก์ชันการอยู่รอด ด้วยวิธีตัวประมาณเบส์โดยวิธีการประมาณที่ไม่ใช้พารามิเตอร์ในที่นี้พิจารณาการแจกแจงก่อนของ F เป็นกระบวนการนิวทรัลทางขวา (Process neutral to the right)

เมื่อ $F(t)$ เป็นฟังก์ชันการแจกแจงสุ่มที่เป็นนิวทรัลทางขวา (neutral to the right) แล้ว $F(t)$ สามารถเขียนได้ในรูปของ

$$F(t) = 1 - e^{-Y_t} \qquad (2.11)$$

---

[3] Ferguson Thomas S. and Eswar G. Phadia, "Bayesian Nonparametric Estimation Based on Censored Data," The Annals of Statistics 7 (1979) : 163-186.

โดยที่ Y_t เป็นกระบวนการซึ่งมีการเพิ่มขึ้นอย่างอิสระ (Independent increment) ซึ่งมีคุณสมบัติดังนี้

1) Y_t เป็นฟังก์ชันไม่ลดลง (Nondecreasing)
2) Y_t ต่อเนื่องทางขวา (Right-continuous)
3) $\lim_{t \to -\infty} Y_t = 0$
4) $\lim_{t \to +\infty} Y_t = \infty$

กำหนดให้ $S(t)$ เป็นฟังก์ชันการอยู่รอดซึ่งนิยามได้ดังนี้

$$S(t) = 1 - F(t) \quad (2.12)$$

ดังนั้น จากสมการ (2.11) และ (2.12) จะได้ว่าการหาค่าคาดหวังของการแจกแจงภายหลังของ F เมื่อกำหนด (Z_i, δ_i) คือ การหาโมเมนต์ที่ 1 ของ Y_t

$$\begin{aligned} E\left[S(t)|(Z_i,\delta_i)\right] &= E\left[(1-F(t))|(Z_i,\delta_i)\right] \\ &= E\left[\left(1-\left(1-e^{-Y_t}\right)\right)|(Z_i,\delta_i)\right] \\ &= E\left[\left(e^{-Y_t}\right)|(Z_i,\delta_i)\right] \end{aligned}$$

นั่นคือ
$$E\left[S(t)|(Z_i,\delta_i)\right] = M_t\left(1|(Z_i,\delta_i)\right) \quad (2.13)$$

โดยที่ $M_t\left(\theta|(Z_i,\delta_i)\right)$ เป็นฟังก์ชันที่ให้โมเมนต์ภายหลัง (Posterior Moment Generating function) ของ Y_t ซึ่งในที่นี้นิยามฟังก์ชันที่ให้โมเมนต์เป็น

$$M_t(\theta) = E\left(e^{-\theta Y_t}\right)$$

กำหนดให้ $u_1, u_2, \ldots, u_k$ คือ ค่าที่แตกต่างกันของ $t_1, t_2, \ldots, t_n$ โดยที่ $u_1 < u_2 < \ldots < u_k$

$\rho_1, \rho_2, \ldots, \rho_k$ คือ จำนวนของค่าสังเกตที่ไม่ถูกตัดทิ้งที่ $u_1, u_2, \ldots, u_k$ ตามลำดับ

$\lambda_1, \lambda_2, \ldots, \lambda_k$ คือ จำนวนของค่าสังเกตที่มีค่าถูกตัดทิ้งที่ $u_1, u_2, \ldots, u_k$ ตามลำดับ

h_j เป็นจำนวนของ t_i ที่มีค่ามากกว่า u_j

$j(t)$ เป็นจำนวนของ u_i ที่มีค่าน้อยกว่าหรือเท่ากับ t

ดังนั้น จะได้ว่า
$$h_j = \sum_{i=j+1}^{k} (\rho_i + \lambda_i)$$

เมื่อ F เป็นฟังก์ชันการแจกแจงสุ่มที่เป็นนิวทรัลทางขวา และให้ $T_1, T_2, \ldots, T_n$ เป็นตัวอย่างสุ่มขนาด n จาก F แล้วจะได้ว่า การแจกแจงภายหลังของ F เมื่อกำหนดข้อมูล (Z_i,δ_i) จะเป็น นิวทรัลทางขวา[4] และมีฟังก์ชันที่ให้โมเมนต์ภายหลังของ Y_t เป็น

$$M_t\big(\theta|(Z_i,\delta_i)\big) = \frac{M_t\big(\theta+h_{j(t)}\big)}{M_t\big(h_{j(t)}\big)} \cdot \prod_{i=1}^{j(t)}\left[\frac{M_{u_i}^-(\theta+h_{i-1})}{M_{u_i}^-(h_{i-1})}\cdot\frac{C_{u_i}(\theta+h_i+\lambda_i,\rho_i)}{C_{u_i}(h_i+\lambda_i,\rho_i)}\cdot\frac{M_{u_i}(h_i)}{M_{u_i}(\theta+h_i)}\right] \qquad (2.14)$$

โดยที่ ถ้า u เป็นจุดที่กำหนดไว้ล่วงหน้าของ Y_t

$$C_u(\alpha,\beta) = \int_0^{\infty} e^{-\alpha s}\left(1-e^{-s}\right)^{\beta} dG_u(s) \qquad (2.15)$$

และ ถ้า u ไม่ได้เป็นจุดที่กำหนดไว้ล่วงหน้าของ Y_t

$$\begin{aligned} C_u(\alpha,\beta) &= \int_0^{\infty} e^{-\alpha s}\left(1-e^{-s}\right)^{\beta-1} dH_u(s) && \text{ถ้า } \beta\ge 1 \qquad (2.16)\\ &= 1 && \text{ถ้า } \beta=0 \end{aligned}$$

เมื่อ S เป็นค่าที่เพิ่มขึ้นของ Y_t ที่ u_t นั่นคือ $S = Y_t - Y_t^-$

$G_u(s)$ เป็นการแจกแจงก่อนของ S ที่ u

$H_u(s)$ เป็นการแจกแจงภายหลังของ S ที่ u เมื่อกำหนด $T=u$

ดังนั้น จากสมการ (2.13) และ (2.14) จะได้ว่า

$$M_t\big(1|(Z_i,\delta_i)\big) = \frac{M_t\big(1+h_{j(t)}\big)}{M_t\big(h_{j(t)}\big)} \cdot \prod_{i=1}^{j(t)}\left[\frac{M_{u_i}^-(1+h_{i-1})}{M_{u_i}^-(h_{i-1})}\cdot\frac{C_{u_i}(1+h_i+\lambda_i,\rho_i)}{C_{u_i}(h_i+\lambda_i,\rho_i)}\cdot\frac{M_{u_i}(h_i)}{M_{u_i}(1+h_i)}\right] \qquad (2.17)$$

---

[4] Doksum, K., "Tailfree and neutral random probabilities and their posterior distributions," The Annals of Probability 2 (1974) : 183-201.

และเมื่อกำหนดให้

$$R_t(h) = \frac{M_t(h+1)}{M_t(h)} \tag{2.18}$$

$$r_u(\alpha,\beta) = C_u(\alpha+1,\beta)/C_u(\alpha,\beta) \tag{2.19}$$

ดังนั้น $E[S(t)|(Z_i,\delta_i)] = M_t(1|(Z_i,\delta_i))$

$$= R_t\left(h_{j(t)}\right)\prod_{i=1}^{j(t)}\frac{R_{u_i}^-(h_{i-1})}{R_{u_i}(h_i)}\cdot r_{u_i}(h_i+\lambda_i,\rho_i) \tag{2.20}$$

กำหนดให้ $Y_t^- = \lim_{s\to t} Y_s$

$M_t^-(\theta)$ เป็นฟังก์ชันที่ให้โมเมนต์ (Moment Generating Function) ของ Y_t^-

$M_t(\theta)$ เป็นฟังก์ชันที่ให้โมเมนต์ (Moment Generating Function) ของ Y_t

โดยที่ ฟังก์ชันที่ให้โมเมนต์ของ Y_t สามารถเขียนให้อยู่ในรูป

$$M_t(\theta) = e^{\gamma(t)\int_0^\infty (e^{-z\theta}-1)dN(z)} \tag{2.21}$$

เมื่อ $\gamma(t)$ เป็นฟังก์ชันต่อเนื่องที่ไม่ลดลง ที่มี $\lim_{t\to-\infty}\gamma(t) = 0$ และ $\lim_{t\to+\infty}\gamma(t) = +\infty$

$N(z)$ เป็นฟังก์ชันต่อเนื่องที่ไม่ลดลงซึ่งเป็นตัววัด (measure) ใดๆ บน $(0,\infty)$

สำหรับในการวิจัยครั้งนี้ พิจารณาการแจกแจงก่อนเป็น 2 แบบ ดังนี้ คือ

3.1) <u>แบบกระบวนการแกมมา</u> (Gamma Process)

โดยทั่วไปเมื่อ Y_t มีการแจกแจงแบบแกมมาแล้ว จะเรียก Y_t เป็นกระบวนการแกมมา ซึ่งสามารถเขียนฟังก์ชันที่ให้โมเมนต์ของ Y_t ได้ในรูป

$$M_t(\theta) = \frac{\tau^{\gamma(t)}}{\Gamma(\gamma(t))}\int_0^\infty e^{-\theta y}e^{\tau y}y^{\gamma(t)-1}dy \tag{2.22}$$

$$= \left(\frac{\tau}{\tau+\theta}\right)^{\gamma(t)}$$

$$= e^{\gamma(t)\int_0^\infty (e^{-z\theta}-1)e^{-\tau z}z^{-1}dz}$$

ในที่นี้ ถ้าให้ $\varphi(\alpha,\beta,N) \quad = \quad \int_0^\infty e^{-\alpha z}\left(1-e^{-z}\right)^\beta dN(z)$ เมื่อ $\beta \geq 1$ (2.23)

$= \quad 1$ เมื่อ $\beta = 0$

ดังนั้น สามารถเขียนทุกฟังก์ชันในสมการ (2.20) ให้อยู่ในรูปของฟังก์ชัน φ ได้ดังนี้

$$C_u(\alpha,\beta) = \varphi(\alpha,\beta,N)/\varphi(0,1,N) \quad \text{เมื่อ } \beta \geq 1 \tag{2.24}$$

$$= 1 \quad \text{เมื่อ } \beta = 0$$

และ จากสมการ (2.18) และ (2.21) จะได้ว่า

$$R_t(h) = e^{-\gamma(t)\varphi(h,1,N)} \tag{2.25}$$

และ จากสมการ (2.19) และ (2.24) จะได้ว่า

$$r_u(\alpha,\beta) = \varphi(\alpha+1,\beta,N)/\varphi(\alpha,\beta,N) \tag{2.26}$$

จาก (2.22) จะกล่าวได้ว่า Y_t มีการแจกแจงแบบแกมมาที่มีพารามิเตอร์เป็น $\gamma(t)$ และ τ โดยที่

$$dN(z) = e^{-z\tau}z^{-1}dz$$

ดังนั้น จากสมการ (2.23) สามารถเขียนฟังก์ชัน $\varphi(\alpha,\beta,N)$ ได้เป็น

$$\varphi(\alpha,\beta,N) = \int_0^\infty e^{-z\alpha}\left(1-e^{-z}\right)^\beta e^{-z\tau}z^{-1}dz$$

ในที่นี้ ฟังก์ชัน N จะหาได้จากพารามิเตอร์ τ ซึ่งสามารถนำไปรวมกับ α ได้ ดังนั้น สำหรับกระบวนการแกมมาแล้ว จะได้

$$\varphi_G(\alpha,\beta) = \int_0^\infty e^{-z\alpha}\left(1-e^{-z}\right)^\beta z^{-1}dz \quad \text{เมื่อ } \beta \geq 1 \tag{2.27}$$

$$= 1 \quad \text{เมื่อ } \beta = 0$$

นั่นคือ $\varphi(\alpha,\beta,N) = \varphi_G(\alpha+\tau,\beta)$ และจาก (2.22) และ (2.27) เมื่อ $\theta = 1$ จะได้

$$\varphi_G(\alpha,1) = \log\left(\frac{\alpha+1}{\alpha}\right) \tag{2.28}$$

และจาก (2.25) และ (2.28) สามารถเขียน $R_t(h)$ ได้เป็น

$$R_t(h) = \left(\frac{h+\tau}{h+\tau+1}\right)^{\gamma(t)} \tag{2.29}$$

ดังนั้น จาก (2.20), (2.26) และ (2.29) จะได้ว่า เมื่อ Y_t เป็นกระบวนการแกมมาด้วยพารามิเตอร์ $\gamma(t)$ และ τ สามารถคำนวณฟังก์ชันการอยู่รอด $\hat{S}_G(t)$ จากการหาค่าคาดหวังภายหลังของ $S(t)$ ดังนี้

$$E\left(S(t)|(Z_i,\delta_i)\right) = \left(\frac{h_{j(t)}+\tau}{h_{j(t)}+\tau+1}\right)^{\gamma(t)} \cdot \prod_{i=1}^{j(t)}\left[\left(\frac{(h_{i-1}+\tau)(h_i+\tau+1)}{(h_{i-1}+\tau+1)(h_i+\tau)}\right)^{\gamma(u_i)} \frac{\varphi_G(h_i+\lambda_i+\tau+1,\rho_i)}{\varphi_G(h_i+\lambda_i+\tau,\rho_i)}\right] \tag{2.30}$$

ในทางปฏิบัติแล้ว จะต้องมีการกำหนดการแจกแจงก่อนการทดลอง (Prior guess) ของ $S(t)$ โดยนำความรู้จากข้อมูลในอดีตมาช่วยในการตัดสินใจ และเพื่อเป็นประโยชน์ในการประมาณค่าพารามิเตอร์ โดยที่การแจกแจงก่อนการทดลองของ $S(t)$ สำหรับกระบวนการแกมมาสามารถคำนวณได้ดังนี้ คือ

$$E\left(S(t)\right) = M_t(1) = \left(\frac{\tau}{\tau+1}\right)^{\gamma(t)}$$

ถ้าให้การแจกแจงก่อนการทดลองของ $S(t)$ คือ $S_0(t)$ แล้ว จะได้

$$S_0(t) = \left(\frac{\tau}{\tau+1}\right)^{\gamma(t)}$$

และเมื่อกำหนดค่าให้กับพารามิเตอร์ τ จะสามารถคำนวณค่าพารามิเตอร์ $\gamma(t)$ สำหรับทุกค่าของ t ได้ดังนี้

$$\gamma(t) \quad = \quad \log\left(S_0(t)\right)/\log\left(\tau/(\tau+1)\right) \tag{2.31}$$

3.2) <u>แบบกระบวนการดีริชเลต์</u> (Dirichlet Process)

ในที่นี้พิจารณากระบวนการดีริชเลต์ ($\mathfrak{D}(\alpha)$) ซึ่งมี α เป็นพารามิเตอร์ โดยที่ α เป็นฟังก์ชันต่อเนื่องทางขวาที่ไม่ลดลง (nondecreasing right-continuous function) บนจำนวนจริง (R) โดยที่มี $\lim_{t\to-\infty}\alpha(t)=0$ และ $\lim_{t\to+\infty}\alpha(t)=\alpha(R)<\infty$

เมื่อ $\mathfrak{D}(\alpha)$ เป็นฟังก์ชันการแจกแจงสุ่มที่เป็นนิวทรัลทางขวาแล้ว จะสามารถเขียนฟังก์ชันที่ให้โมเมนต์ของ Y_t ได้ในรูป

$$M_t(\theta) \quad = \quad \frac{\Gamma\left(\alpha(R)\right)}{\Gamma\left(\alpha(t)\right)\Gamma\left(\alpha(R)-\alpha(t)\right)}\int_0^{\infty} e^{-(\alpha(R)-\alpha(t)+\theta)y}\left(1-e^{\alpha(t)-1}\right)dy \tag{2.32}$$

$$= \quad \frac{\Gamma\left(\alpha(R)\right)\Gamma\left(\alpha(R)-\alpha(t)+\theta\right)}{\Gamma\left(\alpha(R)-\alpha(t)\right)\Gamma\left(\alpha(R)+\theta\right)}$$

$$= \quad e^{\int_0^{\infty}\left(e^{-\theta z}-1\right)dN_t(z)}$$

โดยที่ $$dN_t(z) \quad = \quad \frac{e^{-\alpha(R)z}\left(e^{-\alpha(t)z}-1\right)}{z\left(1-e^{-z}\right)}dz$$

เมื่อ $0<\alpha(t)<\alpha(R)$ และจากฟังก์ชันที่ให้โมเมนต์ของ Y_t ในสมการที่ (2.32) กำหนดให้มีการแจกแจงเป็น ที $\left(\alpha(t),\alpha(R)-\alpha(t)\right)$ ดังนั้น สำหรับ $s<t$ จะได้ $Y_t-Y_s\in$ ที $\left(\alpha(t)-\alpha(s),\alpha(R)-\alpha(t)\right)$

โดยที่สามารถหาค่าของฟังก์ชัน R และ r ได้ นั่นคือ จากสมการ (2.18) และ (2.32) จะได้

$$R_t(h) \quad = \quad \frac{\alpha(R)-\alpha(t)+h}{\alpha(R)+h} \tag{2.33}$$

และสำหรับการหาค่าของ $r_u(a,b)$ จะพิจารณาแยกเป็น 2 กรณี คือ

1. เมื่อ u เป็นจุดที่ถูกกำหนดไว้ล่วงหน้า (prior fixed point) ของ Y_t

ดังนั้น u จะเป็นจุดที่ไม่ต่อเนื่องของ α และถ้าให้ $\Delta(u)=\alpha(u)-\alpha^-(u)$ แล้ว ฟังก์ชันการแจกแจงก่อนของ $Y_u-Y_u^-$ คือ $\text{บ}(\Delta(u),\alpha(R)-\alpha(u))$ และจากสมการ (2.15) และ (2.32) จะได้

$$C_u(a,b) = \frac{\Gamma(\alpha(R)-\alpha(u)+\Delta(u))\Gamma(\Delta(u)+b)\Gamma(\alpha(R)-\alpha(u)+a)}{\Gamma(\Delta(u))\Gamma(\alpha(R)-\alpha(u))\Gamma(\alpha(R)-\alpha(u)+\Delta(u)+a+b)} \qquad (2.34)$$

ดังนั้น $$r_u(a,b) = \frac{\alpha(R)-\alpha(u)+a}{\alpha(R)-\alpha(u)+\Delta(u)+a+b} \qquad (2.35)$$

2. เมื่อ u ไม่ใช่จุดที่กำหนดไว้ล่วงหน้าของ Y_t สามารถคำนวณได้จากสมการ (2.34) โดยพิจารณาที่ $\Delta(u)=0$

ดังนั้นค่าคาดหวังภายหลังของ $S(t)$ เมื่อการแจกแจงก่อนของ F เป็นกระบวนการดีริชเลต์จะคำนวณได้จาก

$$E(S(t)|(Z_i,\delta_i)) = \frac{\alpha(R)-\alpha(t)+h_{j(t)}}{\alpha(R)+n} \cdot \prod_{i=1}^{j(t)} \frac{(\alpha(R)-\alpha^-(u_i)+h_{i-1})(\alpha(R)-\alpha(u_i)+h_i+\lambda_i)}{(\alpha(R)-\alpha(u_i)+h_i)(\alpha(R)-\alpha^-(u_i)+h_i+\lambda_i+\rho_i)} \qquad (2.36)$$

ให้ $$K = \prod_{i=1}^{j(t)} \frac{(\alpha(R)-\alpha^-(u_i)+h_{i-1})(\alpha(R)-\alpha(u_i)+h_i+\lambda_i)}{(\alpha(R)-\alpha(u_i)+h_i)(\alpha(R)-\alpha^-(u_i)+h_i+\lambda_i+\rho_i)}$$

จากสมการ (2.36) พบว่า ถ้า u_i เป็นจุดที่เกิดการตัดทิ้งของข้อมูล (Censoring point) แล้วจะได้ว่า $\rho_i=0$ และทำให้ $h_{i-1}=h_i+\lambda_i$ ดังนั้น

$$K = \prod_{i=1}^{j(t)} \frac{(\alpha(R)-\alpha(u_i)+h_i+\lambda_i)}{(\alpha(R)-\alpha(u_i)+h_i)}$$

และเมื่อ u_i ไม่ใช่จุดที่เกิดการตัดทิ้งของข้อมูล จะได้ว่า $\lambda_i = 0$ และทำให้ $h_{i-1} = h_i + \rho_i$ ดังนั้น

$$K = 1$$

นั่นคือ สามารถคำนวณฟังก์ชันการอยู่รอด $\hat{S}_D(t)$ จากค่าคาดหวังภายหลังของ $S(t)$ เมื่อการแจกแจงก่อนของ F เป็นกระบวนการดีริชเลต์ ดังนี้

$$E\left(S(t)\middle|(Z_i,\delta_i)\right) = \frac{\alpha(R)-\alpha(t)+h_{j(t)}}{\alpha(R)+n}\cdot\prod_{i=1}^{j(t)}\frac{\left(\alpha(R)-\alpha(u_i)+h_i+\lambda_i\right)}{\left(\alpha(R)-\alpha(u_i)+h_i\right)} \qquad (2.37)$$

ในทางปฏิบัติแล้ว จะต้องมีการกำหนดการแจกแจงก่อนการทดลอง (Prior guess) ของ $S(t)$ โดยนำความรู้จากข้อมูลในอดีตมาช่วยในการตัดสินใจ และเพื่อเป็นประโยชน์ในการประมาณค่าพารามิเตอร์ โดยที่การแจกแจงก่อนการทดลองของ $S(t)$ สำหรับกระบวนการดีริชเลต์สามารถคำนวณได้ดังนี้ คือ

$$E\left(S(t)\right) = \frac{\alpha(t)}{\alpha(R)}$$

ถ้าให้การแจกแจงก่อนการทดลองของ $S(t)$ คือ $S_0(t)$ และ $\alpha(R)$ เป็นขนาดตัวอย่างก่อนการทดลอง (prior sample size)[5] แล้ว จะได้ว่า

$$\alpha(t) = S_0(t).\alpha(R) \qquad (2.38)$$

[5] Ferguson, Thomas S, "A Bayesian analysis of some nonparametric problem," The Annals of Statistics 1(1973) : 209-230.