บทที่ 2

แนวคิด ทฤษฎี และสถิติที่เกี่ยวข้อง

แนวคิดและทฤษฎี

ข้อมูลที่ใช้ในการวิจัยครั้งนี้มาจาก 3 การแจกแจง คือ การแจกแจงล็อกนอร์มอล การแจกแจงล็อกโลจิสติก และการแจกแจงไวบูลล์ และกล่าวถึงวิธีการประมาณพารามิเตอร์สำหรับข้อมูลที่ถูกเซ็นเซอร์แบบช่วง โดยใช้วิธีการประมาณพารามิเตอร์ 3 วิธี คือ วิธีการประมาณแบบวิธีภาวะน่าจะเป็นสูงสุด วิธีการประมาณแบบใช้กราฟ และวิธีการประมาณแบบใช้กราฟที่ปรับค่าเอนเอียง

2.1 การแจกแจงที่ใช้ในการวิจัย

สำหรับการวิจัยในครั้งนี้ ศึกษาจากตัวอย่างสุ่มที่มาจากประชากรที่มีการแจกแจงแบบต่างๆ ที่มีลักษณะเบ้ขวา ดังนี้

1. การแจกแจงแบบล็อกนอร์มอล (Lognormal Distribution)

ถ้า X เป็นตัวแปรสุ่มที่มีการแจกแจงแบบล็อกนอร์มอล แล้วฟังก์ชันความหนาแน่นน่าจะเป็น (Probability Density Function : PDF) อยู่ในรูป

$$f(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right], \qquad x>0$$

เมื่อ $\sigma>0$ และ $-\infty<\mu<\infty$ โดยที่ μ เป็นพารามิเตอร์แสดงสเกล (scale parameter) และ σ เป็นพารามิเตอร์แสดงรูปร่าง (shape parameter) ของการแจกแจงแบบล็อกนอร์มอล และมีฟังก์ชันการแจกแจงสะสม (Cumulative Distribution Function : CDF) อยู่ในรูป

$$F(x;\mu,\sigma)=\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$$

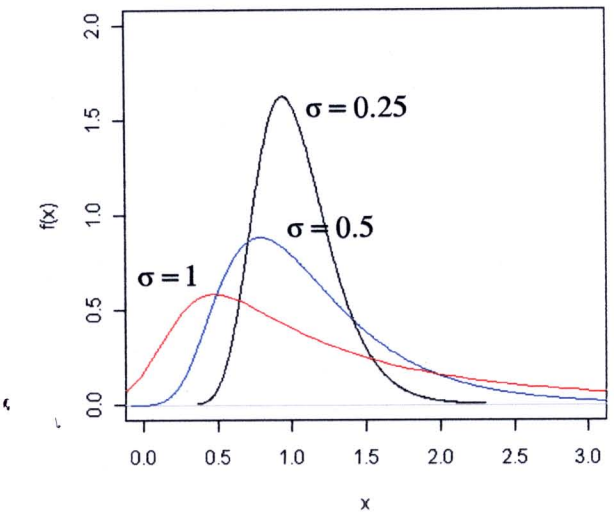
เมื่อ $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[\frac{-t^2}{2}\right]dt$ เป็น Standard Normal CDF

ซึ่งมีค่าเฉลี่ยและค่าความแปรปรวน คือ

$$E(X)=\exp\left(\mu+\frac{1}{2}\sigma^2\right)$$

และ $$Var(X)=\left(\exp(\sigma^2)-1\right)\left(\exp(2\mu+\sigma^2)\right)$$ ตามลำดับ

ภาพที่ 2.1 แสดงฟังก์ชันความหนาแน่นน่าจะเป็นของการแจกแจงแบบล็อกนอร์มอล ที่ $\mu = 0$ และ $\sigma = 0.25, 0.5, 1$



2. การแจกแจงแบบล็อกโลจิสติก (Loglogistic Distribution)

ถ้า X เป็นตัวแปรสุ่มที่มีการแจกแจงแบบล็อกโลจิสติก แล้วฟังก์ชันความหนาแน่นน่าจะเป็น (PDF) อยู่ในรูป

$$f(x;\mu,\sigma)=\frac{1}{\sigma x}\frac{\exp\left[\frac{\ln(x)-\mu}{\sigma}\right]}{\left[1+\exp\left(\frac{\ln(x)-\mu}{\sigma}\right)\right]^2}, \qquad x>0$$

เมื่อ $\sigma > 0$ และ $-\infty < \mu < \infty$ โดยที่ μ เป็นพารามิเตอร์แสดงสเกล (scale parameter) และ σ เป็นพารามิเตอร์แสดงรูปร่าง (shape parameter) ของการแจกแจงแบบล็อกโลจิสติก และมีฟังก์ชันการแจกแจงสะสม (CDF) อยู่ในรูป

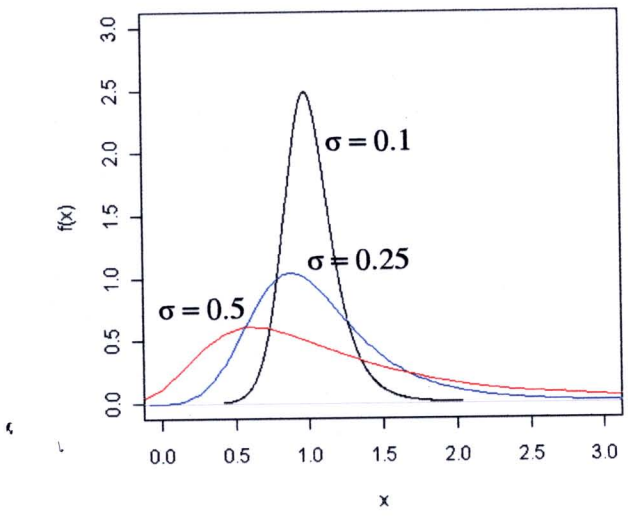
$$F(x;\mu,\sigma)=\frac{\exp\left[\frac{\ln(x)-\mu}{\sigma}\right]}{\left[1+\exp\left(\frac{\ln(x)-\mu}{\sigma}\right)\right]}$$

ซึ่งมีค่าเฉลี่ยและค่าความแปรปรวน คือ

$$E(X)=\exp(\mu)\Gamma(1+\sigma)\Gamma(1-\sigma)$$

และ $$Var(X)=\exp(2\mu)[\Gamma(1+2\sigma)\Gamma(1-2\sigma)-\Gamma^2(1+\sigma)\Gamma^2(1-\sigma)]$$ ตามลำดับ

ภาพที่ 2.2 แสดงฟังก์ชันความหนาแน่นน่าจะเป็นของการแจกแจงแบบล็อกโลจิสติก ที่ $\mu = 0$ และ $\sigma = 0.1, 0.25, 0.5$



3. การแจกแจงแบบไวบูลล์ (Weibull Distribution)

ถ้า X เป็นตัวแปรสุ่มที่มีการแจกแจงแบบไวบูลล์ แล้วฟังก์ชันความหนาแน่นน่าจะเป็น (PDF) อยู่ในรูป

$$f(x;\eta,\beta)=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right], \qquad x>0$$

สำหรับ η และ β เป็นค่าคงที่ เมื่อ $\eta>0$ และ $\beta>0$ โดยที่ η เป็นพารามิเตอร์แสดงสเกล (scale parameter) และ β เป็นพารามิเตอร์แสดงรูปร่าง (shape parameter) ของการแจกแจงแบบไวบูลล์ และมีฟังก์ชันการแจกแจงสะสม (CDF) อยู่ในรูป

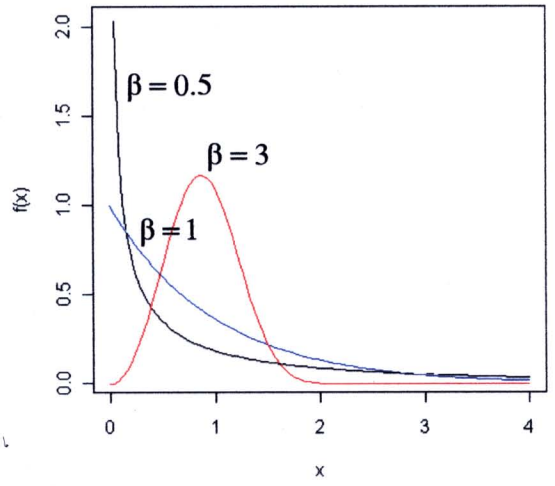
$$F(x;\eta,\beta)=1-\exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right]$$

ซึ่งมีค่าเฉลี่ยและค่าความแปรปรวน คือ

$$E(X)=\eta\Gamma\left(1+\frac{1}{\beta}\right)$$

และ $$Var(X)=\eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\Gamma^2\left(1+\frac{1}{\beta}\right)\right]$$ ตามลำดับ

ภาพที่ 2.3 แสดงฟังก์ชันความหนาแน่นน่าจะเป็นของการแจกแจงแบบไวบูลล์ ที่ $\eta = 1$ และ $\beta = 0.5, 1, 3$



หมายเหตุ ถ้า $\beta = 1$ การแจกแจงแบบไวบูลล์จะมีลักษณะเป็นการแจกแจงแบบเอ็กซ์โปเนนเชียล (Exponential distribution)

4. ข้อมูลที่ถูกเซ็นเซอร์ (Censored Data)

ข้อมูลที่ถูกเซ็นเซอร์ คือ การกำหนดวันเริ่มต้นการศึกษาและวันสิ้นสุดการศึกษา สนใจการเกิดเหตุการณ์ที่สนใจนับจากวันเริ่มต้นการศึกษา โดยที่เมื่อสิ้นสุดการศึกษาแล้วยังไม่เกิดเหตุการณ์ที่สนใจ หรือไม่สามารถบอกได้ว่าเหตุการณ์ที่สนใจเกิดขึ้นเมื่อไร จะทำการตัดข้อมูลทำให้ไม่ทราบค่าจริงของข้อมูลดังกล่าว แต่จะรู้ช่วงของค่าที่เป็นไปได้เท่านั้น (Klein & Moesschberger, 1997; Lawless, 2003; Meeker & Escobar, 1998) อยู่ในรูปของ

$$a \le x \le b \quad ; -\infty \le a < b \le \infty$$

โดยที่ x เป็นค่าของข้อมูลจริง

a และ b เป็นค่าคงที่

ซึ่งสามารถแสดงรูปแบบความสัมพันธ์ได้ในตารางที่ 2.1

ตารางที่ 2.1 แสดงลักษณะของข้อมูลที่ถูกเซ็นเซอร์และฟังก์ชันภาวะน่าจะเป็น

	ชนิดของข้อมูลเซ็นเซอร์	ฟังก์ชันภาวะน่าจะเป็น
ถ้า $a = -\infty$	ข้อมูลเซ็นเซอร์ทางซ้าย (Left Censoring)	$L(\theta;x)=P(x \le b;\theta) = F(b)$
ถ้า $-\infty < a < b < \infty$	ข้อมูลเซ็นเซอร์แบบช่วง (Interval Censoring)	$L(\theta;x)= P(a \le x \le b;\theta) = F(b)-F(a)$
ถ้า $b = \infty$	ข้อมูลเซ็นเซอร์ทางขวา (Right Censoring)	$L(\theta;x)=P(x \ge a;\theta) = 1-F(a)$

2.2 ตัวสถิติที่ใช้ในการวิจัย

การประมาณค่าพารามิเตอร์ที่ใช้ในงานวิจัยครั้งนี้ มีดังนี้

1. การประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation : MLE)

ให้ $X_1,...,X_n$ เป็นตัวอย่างสุ่มขนาด n จากประชากรที่มีฟังก์ชันความหนาแน่นน่าจะเป็น $f(x;\theta)$

ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของตัวอย่างสุ่ม ได้แก่ ฟังก์ชันความหนาแน่นน่าจะเป็นร่วมของ $X_1,...,X_n$ โดยถือว่าเป็นฟังก์ชันของพารามิเตอร์ θ ซึ่งเรามักแทนฟังก์ชันภาวะน่าจะเป็นด้วย L หรือ $L(\theta;x_1,...,x_n)$ หรือ $L(\theta)$ นั่นคือ $L=f(x_1;\theta)f(x_2;\theta)...f(x_n;\theta)$ โดยถือว่า L เป็นฟังก์ชันของ θ

การหาค่าของพารามิเตอร์ θ ที่ทำให้ $L(\theta;x_1,...,x_n)$ มีค่ามากที่สุด (ธีระพร, 2536)

$$\hat{\theta}=\underset{\theta}{\operatorname{argmax}}\,L(\theta;x_1,...,x_n)$$

2. การประมาณด้วยวิธีแบบใช้กราฟ (Graphical Estimation : GE)

ให้ X เป็นตัวแปรสุ่มจากฟังก์ชันการแจกแจง F โดยที่

$$P(X\le x;\mu,\sigma)=F_{\mu,\sigma}(x)$$

และ F มาจากการแจกแจงแบบ location-scale family (Klein & Moesschberger, 1997)

ก็ต่อเมื่อ $$F_{\mu,\sigma}(x)=F_{0,1}\left(\frac{x-\mu}{\sigma}\right)$$

โดยเรียก μ ว่า location parameter เมื่อ $-\infty<\mu<\infty$

σ ว่า scale parameter เมื่อ $\sigma>0$

เนื่องจาก $F_{\mu,\sigma}(x)=F_{0,1}\left(\frac{x-\mu}{\sigma}\right)$ ทำให้

$$x_p=\mu+\sigma F_{0,1}^{-1}(p)$$

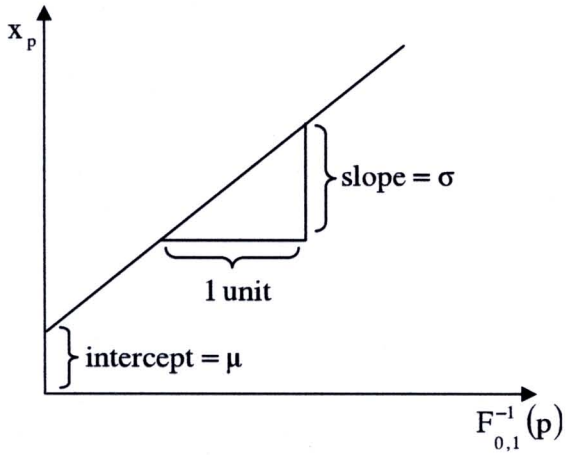
โดยที่ $F_{0,1}^{-1}(p)$ คือ ฟังก์ชันควอนไทล์ (quantile function) ของ $F_{0,1}$

และ x_p คือ ควอนไทล์ที่ p ของ X

จะเห็นว่า ควอนไทล์ของ X ($X\sim F_{\mu,\sigma}$) เป็นฟังก์ชันเชิงเส้นของควอนไทล์จาก $F_{0,1}$

ถ้าเราพล็อตกราฟระหว่าง x_p กับ $F_{0,1}^{-1}(p)$ โดยพล็อต x_p บนแกน y และพล็อต $F_{0,1}^{-1}(p)$ บนแกน x โดยที่ μ และ σ เป็นค่าจุดตัดบนแกน y (intercept) และค่าความชัน (slope) ของเส้นตรง

ภาพที่ 2.4 แสดงการประมาณด้วยวิธีแบบใช้กราฟ



จากความสัมพันธ์ของ x_p และ $F_{0,1}^{-1}(p)$ เราสามารถนำมาใช้ในการประมาณโดยใช้กราฟด้วยการพล็อตกราฟระหว่าง $x_{(i)}$ กับ $F_{0,1}^{-1}(p_{(i)})$ โดยให้ $x_{(i)}$ เป็น sample quantile และ $p_{(i)}$ เป็นลำดับควอนไทล์ของ $x_{(i)}$ ดังนั้น ค่าประมาณของ intercept และ slope จากกราฟดังกล่าว จึงสามารถนำมาใช้ในการประมาณค่า μ และ σ ได้ (Lawless, 2003; Somboonsavatdee & Nair, 2007)

3. การประมาณด้วยวิธีแบบใช้กราฟที่ปรับค่าเอนเอียง (Bias Corrected Graphical Estimation : BCGE)

ให้ **GE** เป็นค่าประมาณพารามิเตอร์ด้วยวิธีแบบใช้กราฟ

ให้ $\mathbf{X}^*$ เป็นตัวอย่างสุ่มจากการประมาณด้วยวิธีแบบใช้กราฟด้วยวิธีบูตสแตรป

วิธีบูตสแตรปเป็นวิธีการประมาณค่าพารามิเตอร์โดยใช้การสุ่มตัวอย่างแบบคืนที่จากตัวอย่างสุ่มชุดเดียวที่มี เพื่อสร้างตัวอย่างชุดใหม่

การหาค่าของพารามิเตอร์ θ^* ทำแบบเดียวกับการประมาณด้วยวิธีแบบใช้กราฟ

จากการแจกแจงแบบ location-scale family จะได้ว่า

$$F_{\mu,\sigma}(x^*) = F_{0,1}\left(\frac{x^* - \mu}{\sigma}\right)$$

$$x_p^* = \mu + \sigma F_{0,1}^{-1}(p)$$

โดยที่ $F_{0,1}^{-1}(p)$ คือ ฟังก์ชันควอนไทล์ (quantile function) ของ $F_{0,1}$

และ x_p^* คือ ควอนไทล์ตัวที่ p ของ X^*

โดยจะทำซ้ำ **B** ครั้ง จะได้ค่าประมาณของพารามิเตอร์ θ^* จำนวน **B** ค่า คือ $\theta_B^* = \hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$

ให้ θ_B^* เป็นค่าประมาณพารามิเตอร์จากวิธีการประมาณแบบใช้กราฟด้วยวิธีบูตสแตรปหาค่าเอนเอียง (Bias Correcting : BC) เพื่อหาค่าประมาณพารามิเตอร์ที่ได้จากการประมาณด้วยวิธีแบบใช้กราฟที่ปรับค่าเอนเอียง (BCGE) (Lawless, 2003; Somboonsavatdee & Nair, 2007)

$$\overline{\theta_B^*} = \frac{\theta_B^*}{n}$$

$$BC = \overline{\theta_B^*} - GE$$

$$BCGE = GE - BC$$

4. ค่าประสิทธิภาพสัมพัทธ์ (Relative efficiency : RE)

กำหนดให้ θ_1 เป็นการประมาณด้วยวิธี MLE

θ_2 เป็นการประมาณด้วยวิธี GE หรือ BCGE

$$RE(\theta_1, \theta_2) = \frac{\overline{MSE}(\theta_2)}{\overline{MSE}(\theta_1)}$$

ถ้า $RE(\theta_1, \theta_2) < 1$ นั่นคือ $\overline{MSE}(\theta_2) < \overline{MSE}(\theta_1)$ กล่าวได้ว่า θ_2 เป็นวิธีประมาณที่มีประสิทธิภาพสัมพัทธ์ดีกว่า θ_1

แต่ถ้า $RE(\theta_1, \theta_2) > 1$ นั่นคือ $\overline{MSE}(\theta_2) > \overline{MSE}(\theta_1)$ กล่าวได้ว่า θ_1 เป็นวิธีประมาณที่มีประสิทธิภาพสัมพัทธ์ดีกว่า θ_2