บทที่ 2

ตัวสถิติและผลงานวิจัยที่เกี่ยวข้อง

การวิเคราะห์ข้อมูลเกี่ยวกับอายุการใช้งาน(life time) หรือช่วงเวลาของความล้มเหลว (failure time) ภายใต้การแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์เป็นเรื่องที่มีความสำคัญเรื่องหนึ่ง ซึ่งมักจะเกี่ยวข้องกับงานทางด้านวิศวกรรมศาสตร์ วิทยาศาสตร์และทางการแพทย์ ข้อมูลลักษณะดังกล่าวมักจะเกิดการตัดปลายเนื่องมาจากข้อกำหนดของเวลาและข้อจำกัดอื่นๆ ในบทนี้จะกล่าวถึงรายละเอียดเกี่ยวกับประเภทของข้อมูลตัดปลาย (type of data censoring) ทฤษฎีบทของการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์ ตัวสถิติทดสอบ และผลงานวิจัยที่เกี่ยวข้องซึ่งมีรายละเอียดต่างๆ ดังนี้

2.1 ประเภทของข้อมูลที่มีค่าถูกตัดปลาย (Type Of Data Censoring)

ก) ข้อมูลถูกตัดปลายประเภทที่ 1 (type I censoring) เป็นลักษณะของข้อมูลที่กำหนดระยะเวลาของข้อมูลที่ถูกตัดปลายไว้ล่วงหน้า เช่น ในการศึกษาเกี่ยวกับอายุการใช้งานของเครื่องจักรชนิดหนึ่งโดยศึกษาในระยะเวลา 8,000 ชั่งโมง เมื่อครบระยะเวลา 8,000 ชั่วโมง ถ้าเครื่องจักรยังทำงานได้จะบันทึกอายุการใช้งานของเครื่องจักร 8,000 ชั่วโมงโดยไม่ทราบอายุการใช้งานจริง เราเรียกข้อมูลที่บันทึกไว้นั้นว่าเป็นข้อมูลตัดปลายประเภทที่ 1

ข) ข้อมูลถูกตัดปลายประเภทที่ 2 (type II censoring) เป็นลักษณะของข้อมูลที่จะต้องกำหนดจำนวนข้อมูลที่ถูกตัดปลายไว้ล่วงหน้า เช่น ในการทดลองเกี่ยวกับประสิทธิภาพของเครื่องใช้ไฟฟ้าชนิดหนึ่งจำนวน 20 เครื่อง แทนที่เราจะทำการทดลองอย่างต่อเนื่องจนกระทั่งเครื่องใช้ไฟฟ้าที่นำมาทดลองทั้ง 20 เครื่องจะเสียหรือใช้งานไม่ได้ เราจะหยุดทำการทดลองเมื่อเครื่องใช้ไฟฟ้าที่นำมาทำการทดลองจะเสียเป็นเครื่องที่ 15 ซึ่งการทดลองแบบนี้จะช่วยประหยัดทั้งเวลาและค่าใช้จ่าย

ค) ข้อมูลถูกตัดปลายแบบสุ่ม (random censoring) เป็นลักษณะของข้อมูลที่คล้ายแบบที่ 1 คือมีการกำหนดระยะเวลาของการทดลองแต่การตัดของข้อมูลนั้นอาจเกิดขึ้นได้ก่อนสิ้นสุดการทดลอง จึงเรียกว่าการตัดแบบสุ่ม ส่วนใหญ่จะพบมากในการทดลองทางการแพทย์ เช่น คนไข้ถอนตัวจากการทดลองก่อนสิ้นสุดการทดลอง จึงทำให้ไม่สามารถบันทึกค่าที่แน่นอนของค่าสังเกตนั้นได้

ในการวิจัยครั้งนี้ผู้วิจัยจะทำการศึกษาทั้งในกรณีที่วิเคราะห์ข้อมูลสมบูรณ์และข้อมูลที่ถูกตัดปลายทางขวาซึ่งข้อมูลที่ถูกตัดปลายทางขวาเป็นข้อมูลถูกตัดปลายประเภทที่ 2 ซึ่งกล่าวถึงข้างต้น

2.2 การแจกแจงที่ใช้ในการวิจัย

1. การแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์

การแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์ มีฟังก์ชันความหนาแน่นอยู่ในรูปแบบ

$$f(x)=\begin{cases}\frac{1}{\theta}\exp(-(\frac{x-\beta}{\theta})), x\geq\beta & ,\theta>0\end{cases}$$

เมื่อ β เป็นพารามิเตอร์ตำแหน่ง (location parameter)

และ θ เป็นพารามิเตอร์สเกล (scale parameter)

ตัวประมาณค่าพารามิเตอร์จะใช้ตัวประมาณภาวะน่าจะเป็นสูงสุด (Maximum likelihood) กล่าวคือ ให้ $X_1,X_2,\ldots X_n$ เป็นตัวอย่างสุ่มของการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์ ซึ่งมีพารามิเตอร์ β และ θ จะหาตัวประมาณภาวะน่าจะเป็นสูงสุดสำหรับพารามิเตอร์แต่ละตัว

เนื่องจาก

$$L(\bar{\theta},\bar{x})=\prod_{i=1}^{n}f(x_i)$$

$$=\theta^{-n}\exp\{\frac{-1}{\theta}\sum_{i=1}^{n}(x_i-\beta)\}$$

ดังนั้น

$$\ln L(\bar{\theta};\bar{x})=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}(x_i-\beta)$$

$$=-n\ln\theta-(\sum_{i=1}^{n}(x_i-\beta))\theta^{-1}$$

เพราะฉะนั้น สมการภาวะน่าจะเป็นอยู่ในรูปและ

$$\frac{\partial}{\partial\theta}\ln L=\frac{-n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}(x_i-\beta)=0$$

$$-n\theta+\sum_{i=1}^{n}(x_i-\beta)=0$$

จะได้ว่า $$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}(x_i)-\hat{\beta} \qquad (1)$$

สำหรับตัวประมาณค่าพารามิเตอร์ตำแหน่ง สามารถหาได้จากตัวสถิติที่เพียงพอ

$$\prod_{i=1}^{n} f(x_i) = \theta^{-n} \exp\{\frac{-1}{\theta}\sum_{i=1}^{n}(x_i - \beta)\} \quad \text{ถ้า } y = \min(x_1, x_2, \ldots x_n) \geq \beta$$

$$= f(\bar{T}(\bar{x}); \bar{\theta}) * h(x)$$

เมื่อ $f(\bar{T}(\bar{x}); \bar{\theta}) = \theta^{-n} \exp\{\frac{-1}{\theta}\sum_{i=1}^{n}(x_i - \beta)\}$,$h(x) = 1$

ดังนั้น $\bar{T}(\bar{x}) = (\min(X_1, X_2, \ldots X_n)'$ เป็นตัวสถิติที่เพียงพอสำหรับ $\bar{\beta}$

ดังนั้น $\beta < (X_1 < X_2 < \ldots < X_n)$ จะได้ว่า $\hat{\beta} = \min\{X_1, X_2, \ldots X_n)$

นั่นคือ $\hat{\beta} = X_1$

แทนค่าใน (1) จะได้ว่า

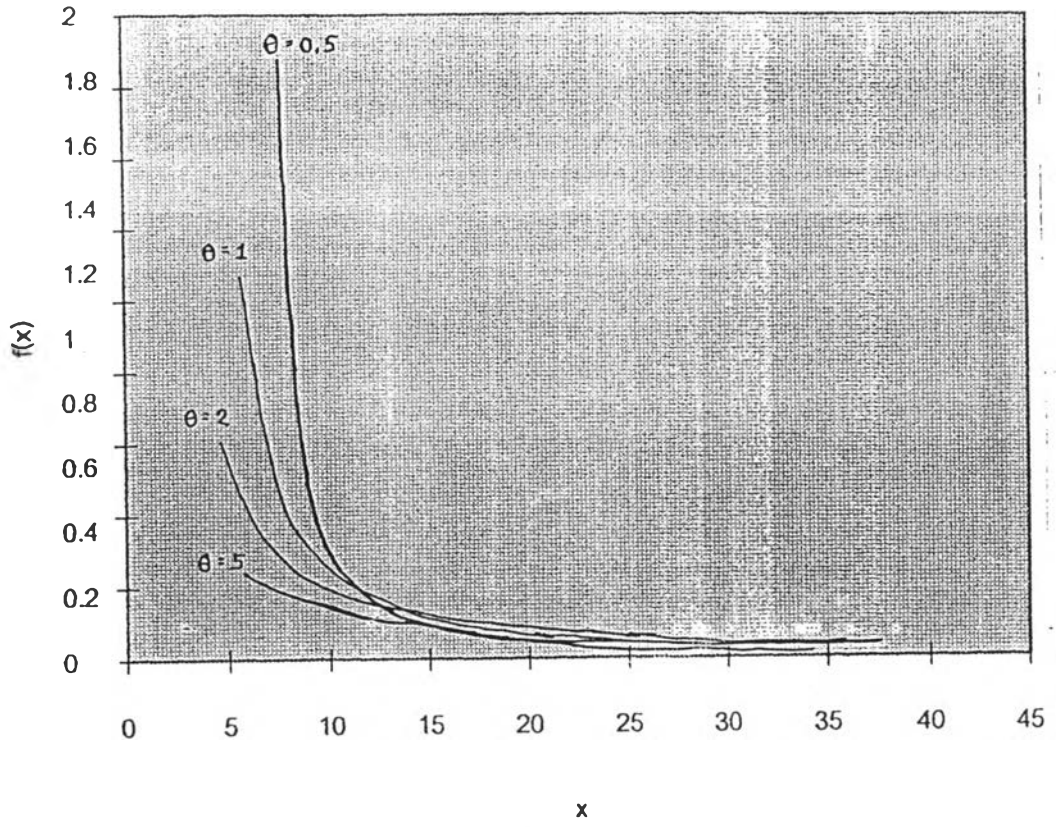
$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i) - \hat{\beta}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i) - X_1$$

ดังนั้นตัวประมาณของพารามิเตอร์ตำแหน่งคือ ตัวสถิติอันดับของข้อมูลที่มีค่าสังเกตน้อยที่สุด และตัวประมาณของพารามิเตอร์สเกลคือ ค่าเฉลี่ยของผลรวมตัวสถิติลำดับ X ตั้งแต่ 1 ถึง n ลบด้วยตัวสถิติอันดับที่หนึ่งของข้อมูลที่มีค่าสังเกตน้อยที่สุด กล่าวคือ

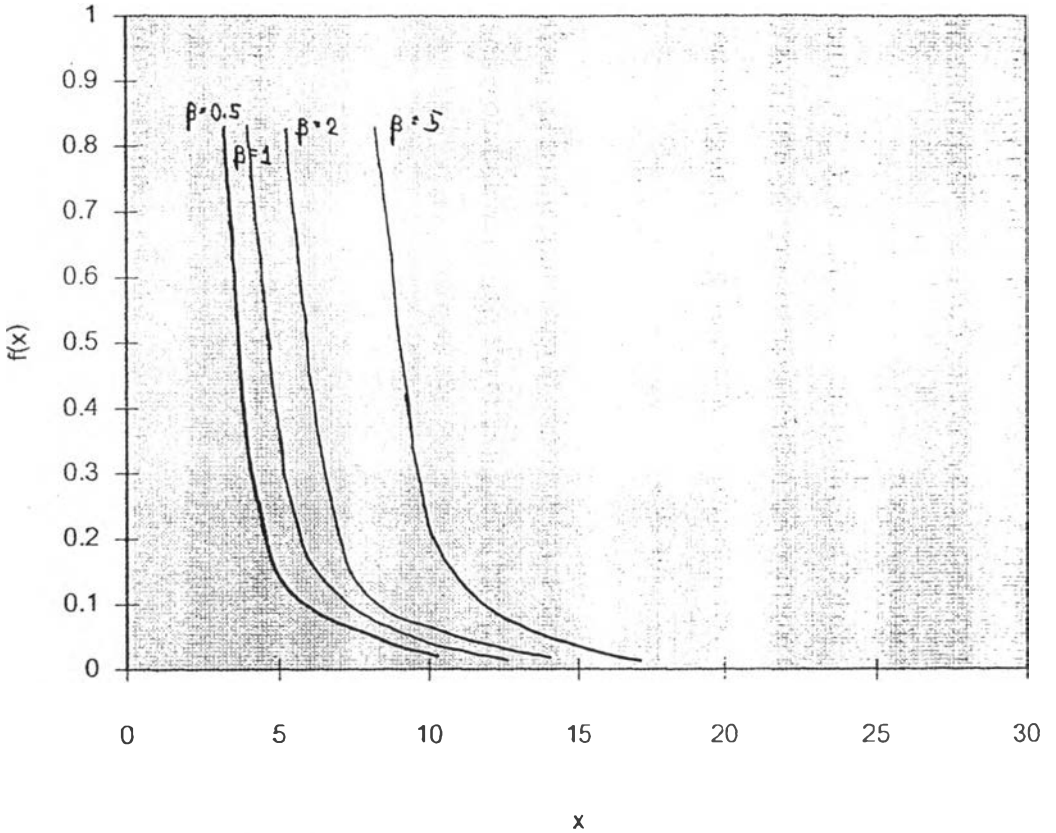
$$\hat{\beta} = X_1 = \min(X_1, X_2, \ldots, X_n)$$

และ $$\hat{\theta} = n^{-1}\sum_{1}^{n} X_i - X_1$$



รูปที่ 2.2.1.1 เส้นโค้งการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์เมื่อ $\beta = 1$ **และ** $\theta = 0.5, 1, 2$ **และ** 5

จากรูปที่ 2.2.1.1 สามารถอธิบายได้ว่าเมื่อพารามิเตอร์สเกล (θ) มีค่าเท่ากับ 0.5 เส้นโค้งจะมีความโค่งมากกว่าค่าพารามิเตอร์สเกลค่าอื่นๆ และเมื่อเพิ่มค่าพารามิเตอร์สเกลเป็น 1, 2 และ 5 เส้นโค้งจะมีความโค่งลดลงและมีแนวโน้มที่ลดต่ำลงเรื่อยๆ ตามค่าของพารามิเตอร์สเกลที่เพิ่มขึ้น



รูปที่ 2.2.1.2 เส้นโค้งการแจกแจงแบบเอชซี้กำลัง 2 พารามิเตอร์เมื่อ $\theta = 1$ และ $\beta = 0.5, 1, 2$ และ 5

จากรูปที่ 2.2.1.2 สามารถอธิบายได้ว่าเมื่อพารามิเตอร์ตำแหน่ง (β) มีค่าเปลี่ยนไป โดยพารามิเตอร์สเกล (θ)มีค่าคงที่ พบว่าลักษณะเส้นโค้งที่ได้มีรูปแบบลักษณะเดียวกัน โดยมีระยะความห่างของแต่ละเส้นโค้งเป็นไปตามค่าพารามิเตอร์ตำแหน่งที่เปลี่ยนไป

2. การแจกแจงแบบแกมมา

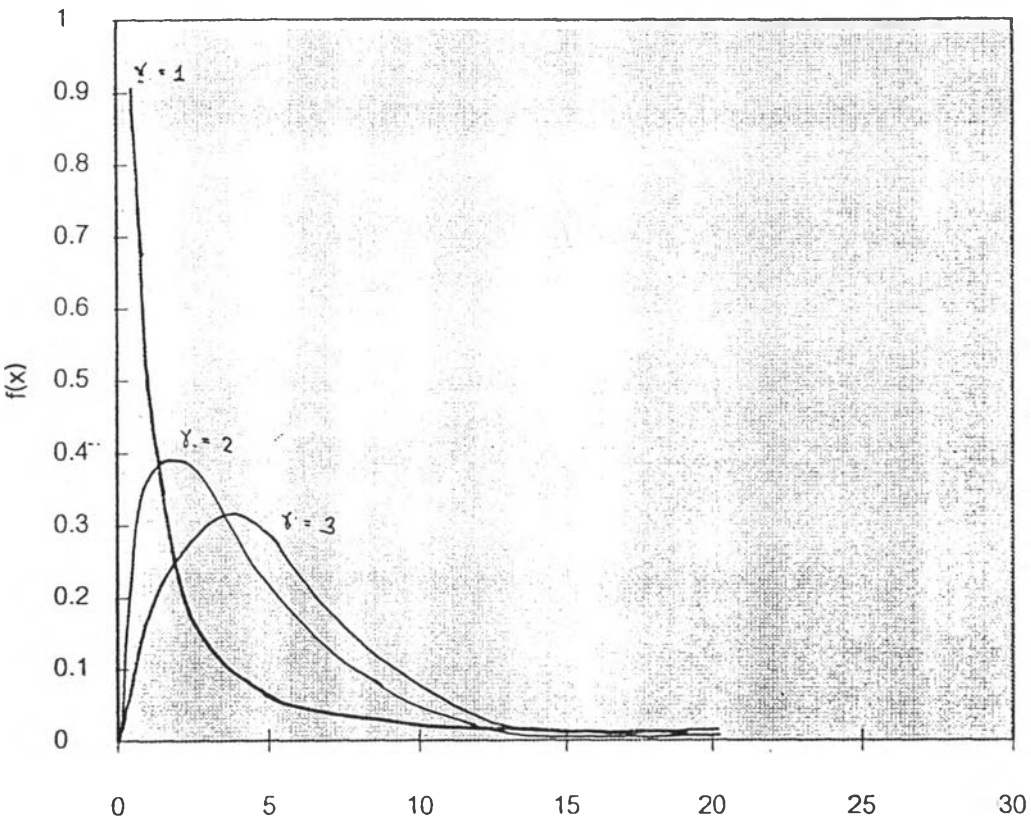
ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = \begin{cases} \dfrac{(x^{\gamma-1}\exp(-x/\lambda))}{\lambda^{\gamma}\,\Gamma(\gamma)} & ,x>0 \quad ,\gamma>0 \quad ,\lambda>0 \\ 0 & \text{อื่นๆ} \end{cases}$$

เมื่อ λ เป็นพารามิเตอร์สเกล (scale parameter) ให้เท่ากับ 1

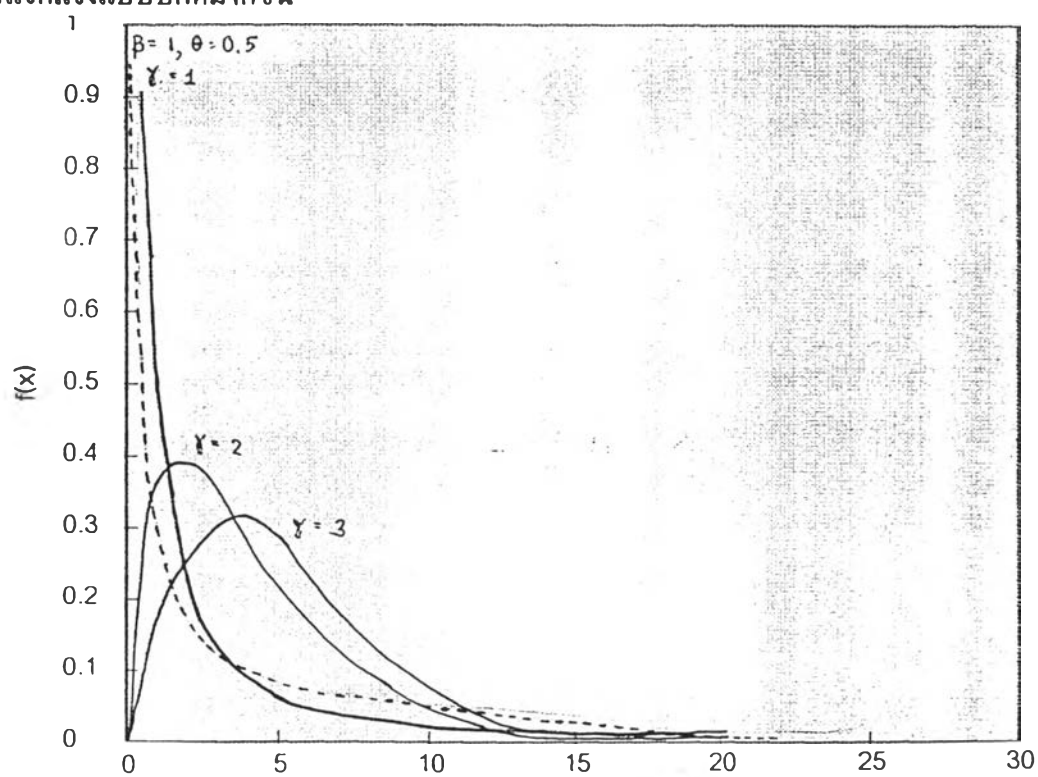
γ เป็นพารามิเตอร์รูปทรง (shape parameter) ให้เท่ากับ 1, 2 และ 3

ซึ่งเขียนรูปกราฟได้ดังนี้



รูปที่ 2.2.2.1 รูปกราฟการแจกจงแบบแกมมา เมื่อ $\lambda = 1$ และ $\gamma = 1, 2$ และ 3 ตามลำดับ

สนใจศึกษาเมื่อค่า $\lambda = 1$ และ $\gamma = 1, 2$ และ 3 โดยมีค่า C.V(X)=100%, 70% และ 50% ตามลำดับ และสาเหตุที่เลือกใช้ค่า C.V(X) เท่ากับ 50% เพราะหากค่า C.V. มีค่าต่ำกว่านี้ลักษณะรูปกราฟจะลู่เข้าสู่การแจกแจงแบบปกติมากขึ้น



รูปที่ 2.2.2.2 รูปกราฟการเปรียบเทียบการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์ $\beta = 1$ และ $\theta = 0.5$ กับการแจกแจงแบบแกมมาเมื่อ $\lambda = 1$ และ $\gamma = 1, 2$ และ 3 ตามลำดับ

3. การแจกแจงแบบไวบูลล์

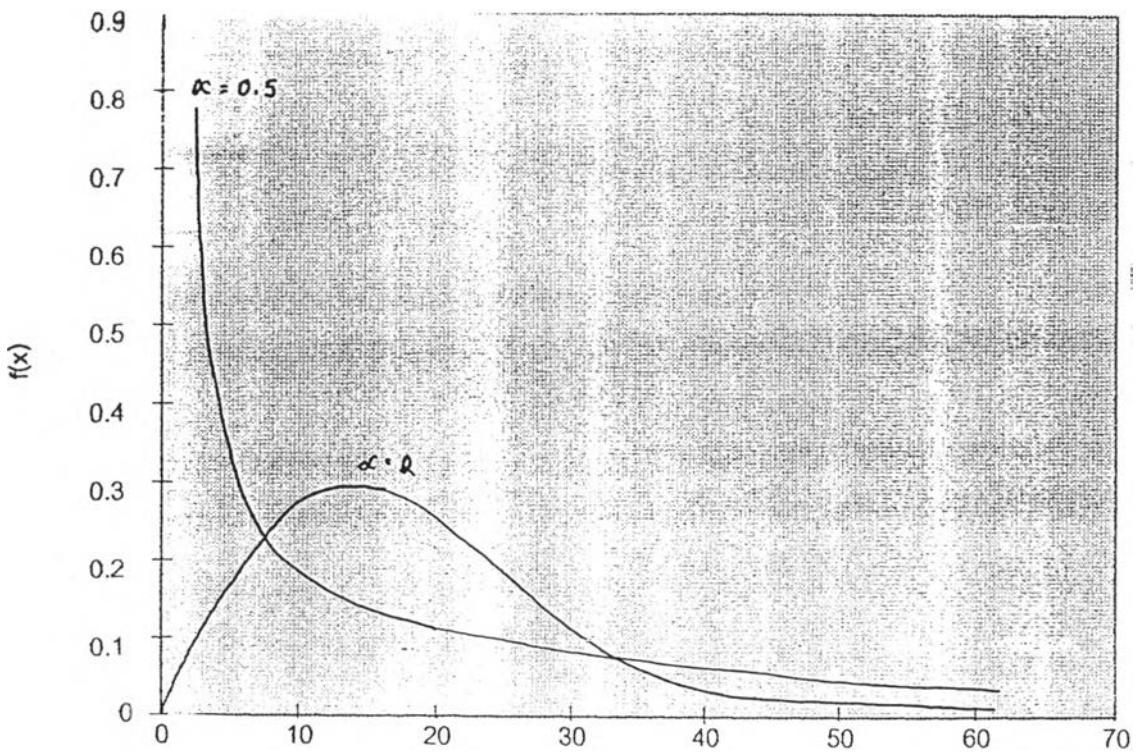
ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x)=\begin{cases}\dfrac{\alpha x^{\alpha-1}\exp[-(x/\beta)^{\alpha}]}{\beta^{\alpha}} & ,0<x<\infty \quad ,\alpha>0 \quad ,\beta>0\\ 0 & \text{อื่นๆ}\end{cases}$$

เมื่อ β เป็นพารามิเตอร์สเกล (scale parameter) ให้เท่ากับ 1

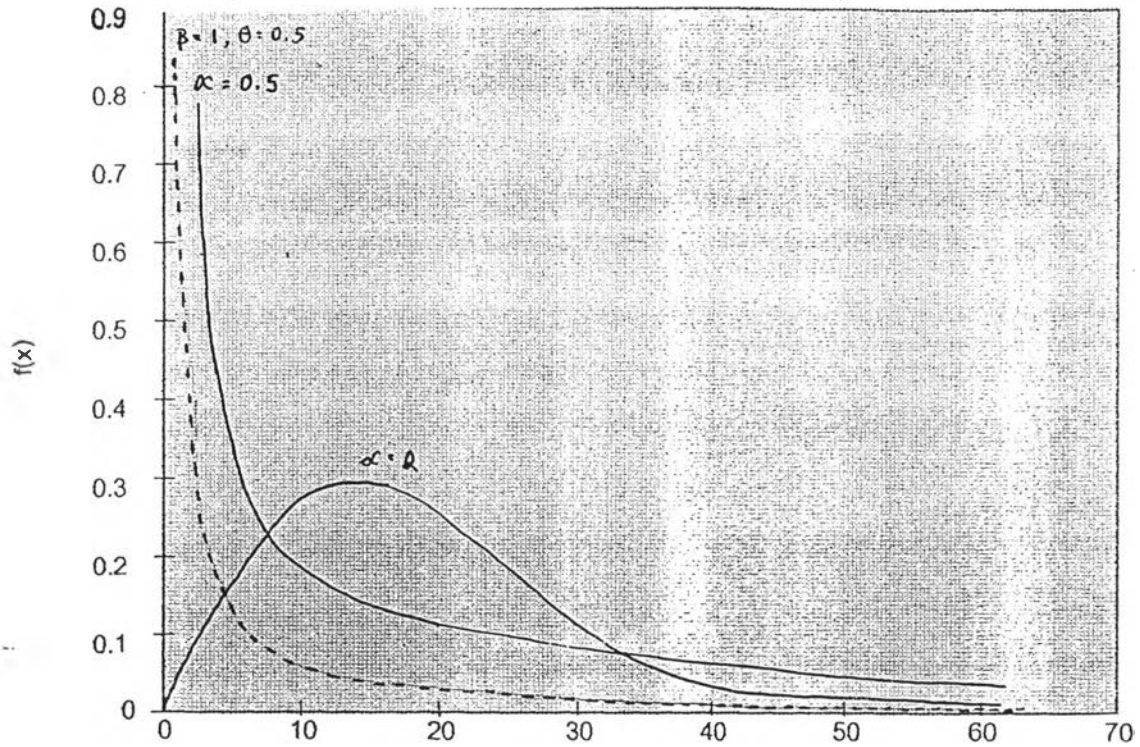
α เป็นพารามิเตอร์รูปทรง (shape parameter) ให้เท่ากับ 0.5 และ 2.0

ซึ่งเขียนรูปกราฟได้ดังนี้



รูปที่ 2.2.3.1 รูปกราฟการแจกแจงแบบไวบูลล์ เมื่อ β = 1 และ α = 0.5 , 2.0 ตามลำดับ

สนใจศึกษาเมื่อค่า β = 1 และ α = 0.5 และ 2.0 โดยมีค่า C.V(X) = 223% และ 52% ตามลำดับ และสาเหตุที่เลือกใช้ค่า C.V(X) เท่ากับ 52% เพราะหากค่า C.V มีค่าต่ำกว่านี้ลักษณะของรูปกราฟจะลู่เข้าสู่การแจกแจงแบบปกติมากขึ้น



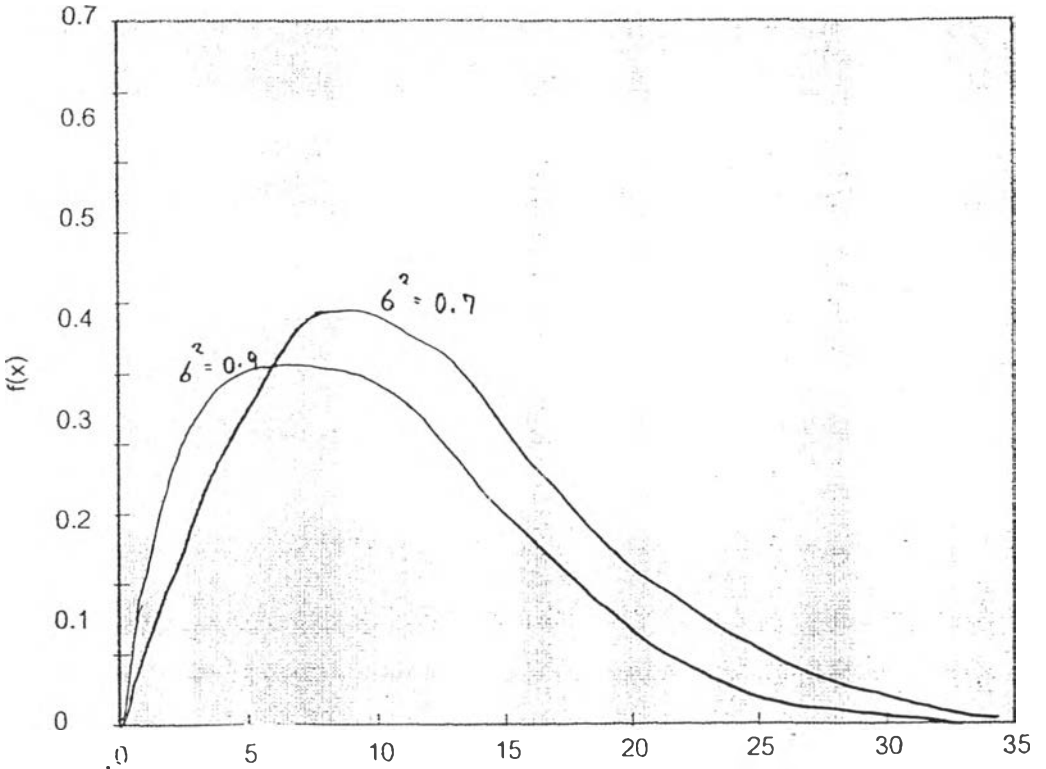
รูปที่ 2.2.3.2 รูปกราฟการเปรียบเทียบการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์ $\beta = 1$ และ $\theta = 0.5$ กับการแจกแจงแบบไวบูลล์เมื่อ $\beta = 1$ และ $\alpha = 0.5$ และ 2.0 ตามลำดับ

4. การแจกแจงแบบลอกนอร์มอล

ฟังก์ชันความหนาแน่นอยู่ในรูป

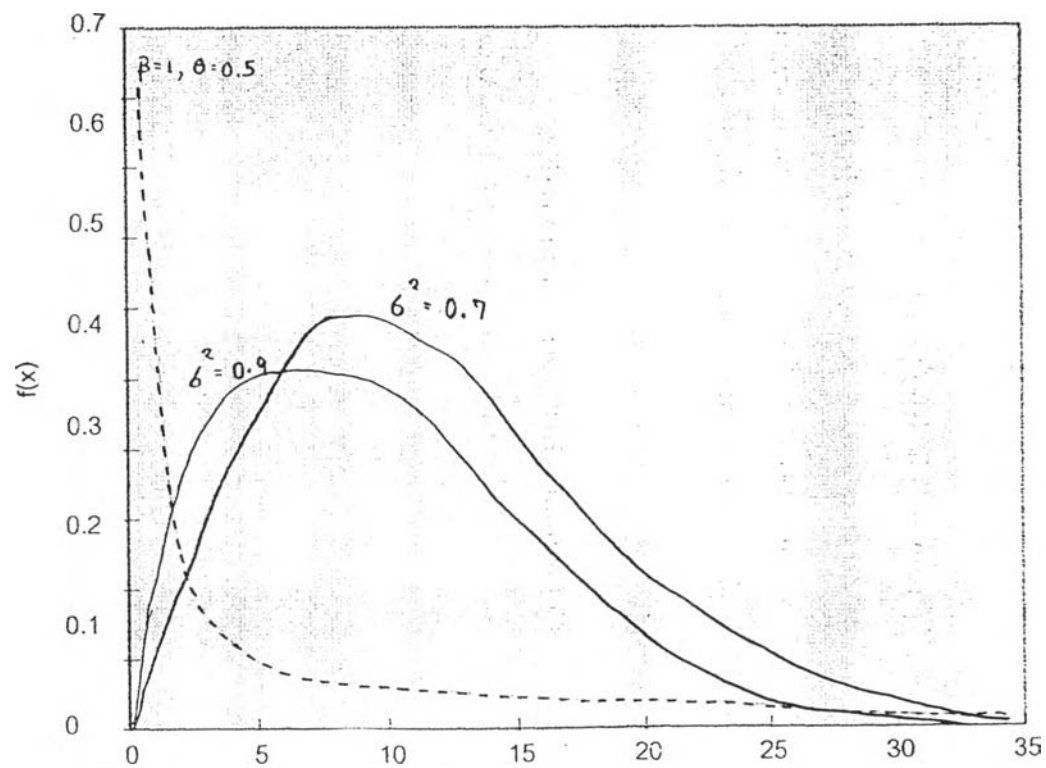
$$f(x) = \begin{cases} \dfrac{\exp[-(\ln x - \mu)^2 / 2\sigma^2]}{x\sqrt{2\pi\sigma^2}} & , x > 0 \quad , \sigma > 0 \quad , -\infty < \mu < \infty \\ 0 & \text{อื่นๆ} \end{cases}$$

เมื่อ μ เป็นค่าเฉลี่ยเท่ากับ 0 และ σ^2 เป็นค่าความแปรปรวนเท่ากับ 0.7 และ 0.9 ซึ่งเขียนรูปกราฟได้ดังนี้



รูปที่ 2.2.4.1 รูปกราฟการแจกแจงแบบลอกนอร์มอล เมื่อ $\mu = 0$ และ $\sigma^2 = 0.7$ และ 0.9 ตามลำดับ

สนใจศึกษาเมื่อ μ เท่ากับ 0 และ σ^2 เท่ากับ 0.7 และ 0.9 มีค่า C.V(X)= 100% และ 120% ตามลำดับโดยค่า C.V มีค่าต่ำกว่านี้รูปกราฟจะลู่เข้าสู่การแจกแจงปกติมากขึ้น



รูปที่ 2.2.4.2 รูปกราฟการเปรียบเทียบการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์ $\beta = 1$ และ $\theta = 0.5$ กับการแจกแจงแบบลอกนอร์มอล เมื่อ $\mu = 0$ และ $\sigma^2 = 0.7$ และ 0.9 ตามลำดับ

2.3 ตัวสถิติทดสอบที่ใช้ในการวิจัย

ก) ตัวสถิติทดสอบ Gini (G)

ในปี ค.ศ. 1978 Gail และ Gastwirth ได้เสนอตัวสถิติทดสอบ G สำหรับการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์กรณีเมื่อไม่ทราบค่าพารามิเตอร์สเกล โดยพัฒนามาจากการศึกษาและนำเสนอของ Dempster และ Kleyle เมื่อปี ค.ศ. 1969 ซึ่งตัวสถิติทดสอบ G อยู่ในรูปแบบ

$$G_n^* = \frac{\sum_{i=1}^{n-1} \{i(n-i)(X_{(i+1)} - X_{(i)})\}}{(n-1)\sum_{i=1}^{n} X_{(i)}} \qquad (1)$$

$X_{(i)}$ เป็นตัวสถิติอันดับจากตัวอย่างสุ่มขนาด n โดยที่ $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$

เมื่อ n มีขนาดใหญ่มากๆ จะประมาณว่าตัวสถิติทดสอบ G จะลู่เข้าสู่การแจกแจงปกติมาตรฐาน กล่าวคือ

$$\frac{G_n^* - E(G_n^*)}{\sqrt{VAR(G_n^*)}} = Z^* \approx N(0,1) \qquad (2)$$

จากการศึกษาของ Hoeffding[1] โดยที่ $E(G_n^*) = 0.5$ สำหรับทุก ๆ ค่าของ n และ $Var(G_n^*) = 1/12(n-1)$ สำหรับค่าวิกฤตของตัวสถิติทดสอบ Gini (G) จะแยกออกเป็น 2 กรณีคือ

1. เมื่อขนาดตัวอย่าง $n \leq 20$ ขอบเขตการยอมรับสมมติฐานสำหรับการทดสอบทางเดียว คือ $|G_n^*| < G_n$(ตาราง)

2. เมื่อขนาดตัวอย่าง $n > 20$ จะต้องปรับค่า G_n^* โดยใช้สมการ (2) ดังนั้นจะปฏิเสธสมมติฐานเมื่อ $|Z^*| > Z_\alpha$

นั่นคือทั้ง 2 กรณีจะปฏิเสธสมมติฐานว่าง H_0: การแจกแจงเป็นแบบเลขชี้กำลัง 2 พารามิเตอร์ เมื่อค่าสถิติทดสอบ G_n^* ที่คำนวณได้มีค่ามากกว่า G_n และ Z^* ที่คำนวณได้มีค่ามากกว่า Z_α จากตาราง ณ ระดับนัยสำคัญที่กำหนด

---

[1] Hoeffding,W.,"A class of statistics with asymptotically normal distribution ,"Ann.Math. Statist.,19,293-325,1949.

ข) ตัวสถิติทดสอบ Lorenz (L)

Gail (ค.ศ. 1977) และ Gail และ Gaswirth(ค.ศ. 1978) ได้ศึกษาหาตัวสถิติทดสอบสำหรับการแจกแจงแบบเลขชี้กำลังที่มี 2 พารามิเตอร์ โดยจะมีอำนาจการทดสอบสูงและสามารถแก้ปัญหาในกรณีที่ข้อมูลตัดปลายได้ดี ซึ่งตัวสถิติทดสอบ L นั้นอยู่ในรูป

$$L_n^*(p) = \frac{\sum_{i=1}^{r=[np]} (X_{(i)} - X_{(1)})}{\sum_{i=1}^{n} (X_{(i)} - X_{(1)})} \quad , \text{โดยที่ } 0 < p < 1 \qquad (3)$$

$X_{(i)}$ เป็นตัวสถิติอันดับจากตัวอย่างสุ่มขนาด n โดยที่ $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ และ $r = [np]$ คือจำนวนเต็มที่ใหญ่ที่สุดที่น้อยกว่าหรือเท่ากับ np เมื่อค่า n มีค่าใหญ่มากๆ ตัวสถิติทดสอบ L จะมีค่าลู่เข้าสู่การแจกแจงปกติมาตรฐาน

$$\frac{L_n^*(p) - \lambda(p)}{\sqrt{Var(L_n(p))}} = Z^* \approx N(0,1) \qquad (4)$$

จากการศึกษาของ Gail,M.H.[2] โดยที่ค่า $\lambda(P)$ คือฟังก์ชันเพิ่มนูน(the convex incresing function) จาก [0,1] ทั่วถึง [0,1] ที่ระดับความน่าจะเป็นเท่ากับ P (การวิจัยครั้งนี้จะใช้ระดับความน่าจะเป็นเท่ากับ 0.5) และ $Var(L_n(P)) = \sigma^2$ และจากทฤษฎีของ Moore จะได้ว่า

$$\lambda(p) = p + (1-p)\ln(1-p)$$

และ $$\sigma^2 = 2q \ln q + p + pq - (p + q \ln q)^2 \quad \text{โดยที่ } q = 1-p$$

สำหรับค่าวิกฤตของตัวสถิติทดสอบ Lorenz (L) จะแยกออกเป็น 2 กรณีคือ

1. เมื่อขนาดตัวอย่าง $n \leq 20$ ขอบเขตการยอมรับสมมติฐานของการทดสอบทางเดียว เมื่อ $|L_n^*(p)| < L_n(p)$ ตาราง

2. เมื่อขนาดตัวอย่าง n > 20 จะต้องปรับค่า $L_n^*(p)$ โดยใช้สมการ (4) ดังนั้นจะปฏิเสธสมมติฐานเมื่อ $|Z^*| > Z_\alpha$

นั่นคือทั้ง 2 กรณีจะปฏิเสธสมมติฐานว่าง H_0: การแจกแจงเป็นแบบเลขชี้กำลัง 2 พารามิเตอร์ เมื่อค่าสถิติทดสอบ $L_n^*(p)$ ที่คำนวณได้มีค่ามากกว่า $L_n(p)$ และ Z^* ที่คำนวณได้มีค่ามากกว่า Z_α จากตาราง ณ ระดับนัยสำคัญที่กำหนด

---

[2]Gail,M.H."A scale-free on the sample Lorenz Currve,"unpublished Ph.D. thesis, Department of statistics,George Washington University.

ค) ตัวสถิติทดสอบ Kolmogorov-Smirnov (K-S)

ในปี ค.ศ. 1976 Barry H. Margorin และ Willi Maurer ได้เสนอตัวสถิติทดสอบ K-S สำหรับการทดสอบการแจกแจงแบบเลขชี้กำลังที่มี 2 พารามิเตอร์ โดยพิจารณาจากค่าความแตกต่างที่สูงที่สุดระหว่างฟังก์ชันการแจกแจงสะสมของตัวอย่างที่จะวิเคราะห์กับฟังก์ชันการแจกแจงสะสมของการแจกแจงที่ต้องการเปรียบเทียบ ตัวสถิติทดสอบ K-S อยู่ในรูปของ

$$K = \max\{D_n^+, D_n^-\}$$

โดยที่

$$D_n^+ = \max_{1\le i\le n}\{\frac{i}{n} - F(X_{(i)}\}$$

$$D_n^- = \max_{1\le i\le n}\{F(X_{(i)}) - \frac{i-1}{n}\}$$

และ $F(X_{(i)})$ คือ ฟังก์ชันการแจกแจงสะสมตามทฤษฎี (i = 1,2,... n) และ n คือจำนวนข้อมูล เราจะปฏิเสธสมมติฐานว่างเมื่อ $K > K_\alpha^*$ จากตาราง Kolmogorov-Smirnov สำหรับการแจกแจงใดๆ

ง) ตัวสถิติทดสอบ Anderson-Darling (A-D)

ตัวสถิติทดสอบ A-D คิดขึ้นโดย Anderson Darling(1953) ซึ่งคำนวณได้ดังนี้

$$A^2 = \frac{-1}{n}[\sum_{i=1}^{n}(2i-1)\{\ln(z_i) + \ln(1 - z_{n+1-i})\}] - n$$

เมื่อ n แทนขนาดตัวอย่าง

z_i แทนความน่าจะเป็นสะสมของการแจกแจงแบบเลขชี้กำลัง 2 พารามิเตอร์

สำหรับค่าวิกฤติของตัวสถิติทดสอบ A-D โดยเปรียบเทียบกับตาราง Anderson-darling ของ M.A. Stephen จะปฏิเสธสมมติฐานว่างเมื่อ A^2 ที่คำนวณได้มีค่ามากกว่าค่าของ percentage points ของตาราง A-D ณ ระดับนัยสำคัญที่กำหนด