การเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนลของตัวประมาณค่าเฉลี่ยถ่วงน้ำหนักในกรณีข้อมูลมีค่าผิดปกติ

Bandwidth Selection for Kernel Density Estimation of Weighted Mean Estimator in Case of Data with Outliers

คณิศา โชติจันทึก
Kanisa Chodjuntug
ภาควิชาคณิตศาสตร์ สถิติและคอมพิวเตอร์ มหาวิทยาลัยอุบลราชธานี 34190

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบประสิทธิภาพวิธีการเลือกแบนวิดจ์ 3 วิธี ได้แก่ วิธี least square cross-validation วิธี maximum likelihood cross-validation และวิธี plug-in สำหรับการประมาณความหนาแน่นแบบเคอร์เนลของค่าเฉลี่ยถ่วงน้ำหนักและค่าเฉลี่ยของตัวอย่าง ในกรณีข้อมูลมีค่าผิดปกติ กำหนดสัดส่วนข้อมูลผิดปกติเท่ากับ 0.05, 0.10, 0.20, 0.30 และ 0.40 โดยใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ ผลการวิจัยสรุปได้ดังนี้ วิธี plug-in มีประสิทธิภาพดีที่สุด เมื่อสัดส่วนข้อมูลผิดปกติเท่ากับ 0.05 และ 0.10 และวิธี least square cross-validation มีประสิทธิภาพดีที่สุด เมื่อสัดส่วนข้อมูลผิดปกติเท่ากับ 0.20, 0.30 และ 0.40

คำสำคัญ : การเลือกแบนวิดจ์ การประมาณความหนาแน่นแบบเคอร์เนล ค่าผิดปกติ

Abstract

The objective of this research is to compare three bandwidth selection methods, the least square cross-validation, maximum likelihood cross-validation and plug-in methods, for kernel density estimation of weighted mean estimator and the sample mean in case of data with outliers. The proportion of outliers in the study equal to 0.05, 0.10, 0.20, 0.30 and 0.40 and the mean square error is used on the basis of the performance comparison. It is found from the research that plug-in method has the best performance when proportion of outliers equal to 0.05 and 0.10 and least square cross-validation has the best performance when proportion of outliers equal to 0.20, 0.30 and 0.40.

Keywords: bandwidth selection, kernel density estimation, outlier

E-mail address : kanisa.c@ubu.ac.th

1. บทนำ

ค่าเฉลี่ย (Mean) มีความสำคัญในทางสถิติเนื่องจากเป็นค่ากลางของข้อมูลที่นิยมใช้ และเป็นตัวประมาณค่าพารามิเตอร์แบบจุดของการแจกแจงหลายชนิด เช่น การแจกแจงแบบปัวซง (Poisson distribution) การแจกแจงแบบปกติ (Normal distribution) เป็นต้น ค่าเฉลี่ยเป็นตัวแทนของข้อมูลเนื่องจากมีคุณสมบัติไม่เอนเอียง มีความคงเส้นคงวาและความแปรปรวนต่ำที่สุด แต่ค่าเฉลี่ยมีข้อจำกัดในการใช้หากข้อมูลมีค่าผิดปกติ (Outlier) ส่งผลให้ค่าที่คำนวณได้มีความคลาดเคลื่อนเกิดขึ้น ในทางปฏิบัติการเก็บข้อมูลมักพบค่าผิดปกติ อาจก่อให้เกิดความเสียหายหากนำค่าเฉลี่ยของตัวอย่าง $(\overline{x})$ ไปประมาณค่าพารามิเตอร์ เนื่องจากค่าประมาณที่ได้จะมีความน่าเชื่อถือลดลง จึงมีการแก้ปัญหาโดยตัดค่าผิดปกติหรือลดน้ำหนักค่าผิดปกติให้น้อยกว่าค่าสังเกตตัวอื่น แต่การแก้ปัญหาโดยตัดค่าผิดปกติออกไปทันทีนั้นอาจไม่เหมาะสม เพราะในบางครั้งข้อมูลที่มีค่าผิดปกติอาจบ่งบอกให้ทราบสารสนเทศบางประการเกี่ยวกับข้อมูล ดังนั้นจึงควรตรวจสอบและค้นหาสาเหตุการมีค่าผิดปกติของข้อมูลนั้นก่อน หากทราบว่าสาเหตุที่เกิดขึ้นเนื่องจากความผิดพลาดในการวัดหรือการบันทึกสามารถตัดข้อมูลที่เป็นค่าผิดปกติออกไป หากไม่สามารถหาสาเหตุและอธิบายถึงความผิดปกติของข้อมูลได้ควรหาวิธีที่เหมาะสมในการประมาณค่าเมื่อข้อมูลมีค่าผิดปกติ [1] มีผู้เสนอตัวประมาณค่าที่สร้างจากแนวคิดการลดน้ำหนักค่าสังเกตที่เป็นค่าผิดปกติเพื่อลดอิทธิพลของค่าผิดปกติให้มีความเหมาะสม เช่น Huber [2] และ Hampel [3] ได้เสนอตัวประมาณค่า Huber estimator Huber-type skipped mean Three-part redescending estimator เป็นต้น ซึ่งตัวประมาณเหล่านี้มีแนวคิดการลดน้ำหนักค่าผิดปกติให้น้อยกว่าค่าสังเกตตัวอื่น โดยกำหนดฟังก์ชันที่ใช้เป็นตัวถ่วงน้ำหนักค่าสังเกตในการประมาณค่า ต่อมา กนกกาญ [4] ได้เสนอตัวประมาณค่าเฉลี่ยถ่วงน้ำหนักด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล $(\overline{x}_K)$ ดังสมการ (1)

$$\overline{x}_K = \sum_{i=1}^{n} \frac{x_i \hat{f}(x)}{\sum_{i=1}^{n} \hat{f}(x)} \tag{1}$$

เมื่อ $\hat{f}(x)$ คือ ค่าประมาณความหนาแน่นแบบเคอร์เนล

การประมาณความหนาแน่นแบบเคอร์เนล (Kernel density estimation) เป็นการประมาณฟังก์ชันความหนาแน่นน่าจะเป็น $f(x)$ ทางด้านนอนพาราเมตริก โดยหาฟังก์ชันโครงสร้างข้อมูลทางคณิตศาสตร์ที่สามารถปรับเปลี่ยนรูปแบบให้เหมาะสมกับข้อมูล [5] สามารถหาตัวประมาณความหนาแน่นแบบเคอร์เนลได้ดังสมการ (2)

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) \tag{2}$$

โดยที่ h คือ ค่าแบนวิดจ์ (Bandwidth)
n คือ จำนวนข้อมูลทั้งหมด
x คือ ตัวแปรที่ไม่ทราบค่า
X_i คือ ตัวแปรสุ่มที่ $i; i = 1, 2, \ldots, n$

$K(x)$ คือ ฟังก์ชันเคอร์เนล (Kernel function)

ฟังก์ชันเคอร์เนลมีหลายรูปแบบซึ่ง Wolfgang [6] ได้สรุปฟังก์ชันเคอร์เนลดังนี้ ฟังก์ชันเคอร์เนลแบบอีพานีชนิคอฟ (Epanechnikov) ฟังก์ชันเคอร์เนลแบบควอติ (Quartic) ฟังก์ชันเคอร์เนลไตรเวท (Triweight) ฟังก์ชันเคอร์เนลแบบเกาส์เซียน (Gaussian) ฟังก์ชันเคอร์เนลแบบสามเหลี่ยม (Triangular) ฟังก์ชันเคอร์เนลแบบยูนิฟอร์ม (Uniform) ฟังก์ชันเคอร์เนลแบบคอซายอัส (Cosinus) แต่จากการศึกษาของ Ahmad and Mugdadi [7] พบว่าในกรณีที่ตัวแปรสุ่มเป็นตัวแปรสุ่มที่เป็นอิสระและมีการแจกแจงเดียวกัน ฟังก์ชันเคอร์เนลที่แตกต่างกันไม่ใช่สิ่งที่สำคัญในการประมาณความหนาแน่น $\hat{f}(x)$ แต่สิ่งที่สำคัญคือ การเลือกแบนวิดจ์ที่เหมาะสม

งานวิจัยนี้จึงสนใจศึกษาวิธีการเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนลของค่าเฉลี่ยถ่วงน้ำหนักคือ วิธีการเลือกแบนวิดจ์แบบ least square cross-validation วิธี maximum likelihood cross- validation และวิธี plug-in ใช้ฟังก์ชันเคอร์เนลแบบเกาส์เซียนเนื่องจากฟังก์ชันเคอร์เนลที่แตกต่างกันไม่ใช่สิ่งที่สำคัญในการประมาณความหนาแน่น $\hat{f}(x)$ [7]

2. วิธีการดำเนินงานวิจัย

การดำเนินการวิจัยมีขั้นตอนดังนี้

2.1 จำลองข้อมูลที่มีการแจกแจงปกติ $N(\mu, \sigma^2)$

2.1.1 ข้อมูลปกติ กำหนดพารามิเตอร์ $\mu = 0, \sigma = 1$

2.1.2 ข้อมูลผิดปกติ กำหนดพารามิเตอร์ $\mu = 4, \sigma = 1$

2.1.3 กำหนดสัดส่วนข้อมูลผิดปกติ (p) เท่ากับ 0, 0.05, 0.10, 0.20, 0.30 และ 0.40

2.2 กำหนดขนาดตัวอย่างทั้งหมด (n) เท่ากับ 25, 40 และ 100

2.3 ขั้นตอนการหาค่าประมาณค่าเฉลี่ยถ่วงน้ำหนัก

2.3.1 กำหนดฟังก์ชันเคอร์เนลแบบเกาส์เซียน (Gaussian Kernel) ดังสมการ (3)

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \tag{3}$$

เมื่อ $u = \frac{x - X_i}{h}$

2.3.2 วิธีการเลือกแบนวิดจ์

1) วิธี Least Square Cross-Validation (LSCV)

Rudemo [8] ได้เสนอการเลือกแบนวิดจ์ด้วยวิธี LSCV ซึ่งเป็นวิธีที่นิยมใช้ ต่อมา Bowman [9] ได้ปรับรูปแบบวิธี LSCV ให้อยู่ในรูปอย่างง่ายดังสมการ (4)

$$LSCV(h) = \int_{-\infty}^{\infty} \hat{f}^2(x)dx - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K(X_i - X_j) \tag{4}$$

เมื่อฟังก์ชันเคอร์เนลเป็นแบบเกาส์เซียนสามารถเขียนสมการ (4) ใหม่ดังสมการ (5)

$$\mathrm{LSCV}(h) = \frac{1}{2n^2 h\sqrt{\pi}}\sum_{i=1}^{n}\sum_{j=i}^{n}\exp-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2 - \frac{1}{n(n-1)h\sqrt{2\pi}}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\exp-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2 \quad (5)$$

สามารถเลือกแบนวิดจ์ โดย $h_{lscv} = \arg\min \mathrm{LSCV}(h)$ เมื่อ $\arg\min \mathrm{LSCV}(h)$ คือ ค่าของ h ที่ทำให้ $\mathrm{LSCV}(h)$ มีค่าน้อยที่สุด

2) วิธี Maximum likelihood Cross- Validation (MLCV)

Duin [10] ได้เสนอการเลือกแบนวิดจ์ด้วยวิธี MLCV ดังสมการ (6)

$$\mathrm{MLCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{j\neq i}^{n}K\left(\frac{X_i - X_j}{h}\right)\right) - \log\left((n-1)h\right) \quad (6)$$

สามารถเลือกแบนวิดจ์ โดย $h_{mlcv} = \arg\max \mathrm{MLCV}(h)$ เมื่อ $\arg\max \mathrm{MLCV}(h)$ คือ ค่าของ h ที่ทำให้ $\mathrm{MLCV}(h)$ มีค่ามากที่สุด

3) วิธี Plug-in

Woodroofe [11] ได้เสนอการเลือกแบนวิดจ์ด้วยวิธี plug-in เป็นครั้งแรก โดยเสนอสมการสำหรับประมาณค่าแบนวิดจ์ที่เหมาะสม (Optimal Bandwidth; h_{opt})

$$h_{opt} = \left(\frac{R(K)}{\mu_2^2(K)R(f'')}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \quad (7)$$

เมื่อ $R(K) = \int_{-\infty}^{\infty} K(x)^2\,dx$ และ $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x)\,dx$

โดยที่ $K(x)$ คือ ฟังก์ชันเคอร์เนล

n คือ จำนวนข้อมูลทั้งหมด

f'' คือ อนุพันธ์อันดับสองของ $f(x)$

เมื่อฟังก์ชันเคอร์เนลเป็นแบบเกาส์เซียนสามารถเขียนสมการ (7) ใหม่ [12] ดังสมการ (8)

$$h_{opt} = \sigma\left(\frac{4}{3n}\right)^{\frac{1}{5}} \quad (8)$$

โดยที่ σ คือ ค่าเบี่ยงเบนมาตรฐาน

2.3.3 คำนวณหาตัวประมาณค่าเฉลี่ยถ่วงน้ำหนักด้วยเคอร์เนล $(\overline{x}_K)$

2.4 เปรียบเทียบวิธีการเลือกแบนวิดจ์

การเปรียบเทียบวิธีการเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนลของค่าเฉลี่ยถ่วงน้ำหนัก โดยพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) คำนวณจากสมการ (9)

$$MSE = \frac{\sum_{i=1}^{1,000} \left(\hat{\theta} - \theta\right)^2}{1,000} \quad (9)$$

โดยที่ $\hat{\theta}$ คือ ค่าประมาณของพารามิเตอร์

θ คือ ค่าพารามิเตอร์

ในงานวิจัยนี้กำหนดให้ ทำซ้ำ 1,000 รอบในแต่ละสถานการณ์ ถ้าแบนวิดจ์ที่คำนวณได้จากวิธีใดมีผลทำให้การประมาณค่าเฉลี่ยแบบถ่วงน้ำหนักด้วยเคอร์เนล มี MSE ต่ำสุด จะมีประสิทธิภาพดีที่สุด

3. ผลการทดลอง

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบวิธีการเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนลของค่าเฉลี่ยถ่วงน้ำหนัก ได้แก่ วิธี LSCV MLCV และ Plug-in และเปรียบเทียบตัวประมาณค่าเฉลี่ยแบบถ่วงน้ำหนักกับค่าเฉลี่ยตัวอย่าง เมื่อข้อมูลมีค่าผิดปกติ โดยพิจารณาค่า MSE ดังตารางที่ 1 และรูปที่ 1 สามารถสรุปได้ดังนี้

3.1 สัดส่วนข้อมูลผิดปกติเพิ่มขึ้นมีผลให้ค่า MSE มีแนวโน้มเพิ่มขึ้น เมื่อพิจารณาวิธีการประมาณค่าเฉลี่ย พบว่า เมื่อ p เท่ากับ 0 ค่า MSE ของ $\overline{x}$ มีค่าต่ำที่สุด ทุกขนาดตัวอย่าง เมื่อ p เท่ากับ 0.05 และ 0.01 ค่า MSE ของค่าเฉลี่ยแบบถ่วงน้ำหนักด้วยวิธี Plug-in มีค่าต่ำที่สุด ทุกขนาดตัวอย่าง ยกเว้น p เท่ากับ 0.05 ที่ขนาดตัวอย่างเท่ากับ 25 และเมื่อ p เท่ากับ 0.20, 0.30, 0.40 ค่า MSE ของค่าเฉลี่ยแบบถ่วงน้ำหนักด้วยวิธี LSCV มีค่าต่ำที่สุด ทุกขนาดตัวอย่าง ยกเว้น p เท่ากับ 0.40 ที่ขนาดตัวอย่างเท่ากับ 25

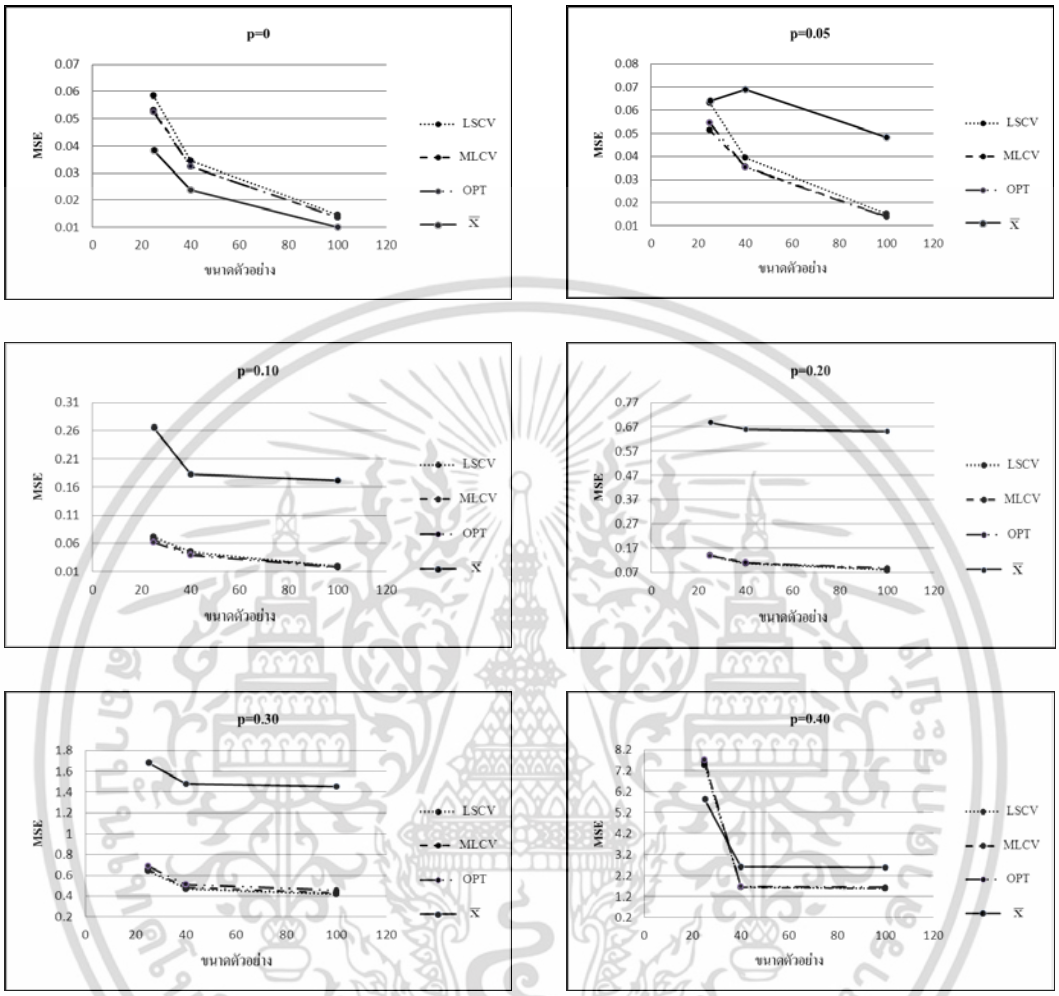
3.2 วิธีการประมาณค่าเฉลี่ยแบบถ่วงน้ำหนักด้วยวิธีการเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนลทั้ง 3 วิธี มีค่า MSE ใกล้เคียงกัน และมีค่าต่ำกว่า $\overline{x}$ ทุกกรณี ยกเว้น p เท่ากับ 0

3.3 ขนาดตัวอย่างเพิ่มขึ้นมีผลให้ค่า MSE มีแนวโน้มลดลง ในทุกวิธีการประมาณค่าฉลี่ย ยกเว้น $\overline{x}$ ที่ p เท่ากับ 0.05

ตารางที่ 1. ค่า MSE ของตัวประมาณค่าเฉลี่ย โดยจำแนกตามขนาดตัวอย่างและสัดส่วน

n	p	$\bar{x}$	$\bar{x}_K$		
			LSCV	Plug-in	MLCV
25	0.00	0.0383*	0.0584	0.0523	0.0532
	0.05	0.0639	0.0632	0.0544	0.0513*
	0.10	0.2650	0.0717	0.0618*	0.0669
	0.20	0.6877	0.1363*	0.1367	0.1411
	0.30	1.6850	0.6440*	0.6853	0.6534
	0.40	5.8313*	7.4611	7.7074	7.6317
40	0.00	0.0236*	0.0345	0.0324	0.0324
	0.05	0.0689	0.0394	0.0351*	0.0355
	0.10	0.1823	0.0448	0.0385*	0.0408
	0.20	0.6607	0.1075*	0.1097	0.1090
	0.30	1.4777	0.4683*	0.5044	0.4796
	0.40	2.5940	1.6307*	1.6751	1.6349
100	0.00	0.0100*	0.0147	0.0138	0.0137
	0.05	0.0484	0.0151	0.0137*	0.0139
	0.10	0.1715	0.0188	0.0175*	0.0180
	0.20	0.6526	0.0793*	0.0867	0.0819
	0.30	1.4508	0.4182*	0.4567	0.4284
	0.40	2.5703	1.5689*	1.6256	1.5828

* หมายถึง ค่า MSE ต่ำสุด



รูปที่ 1. ความสัมพันธ์ระหว่างขนาดตัวอย่างและ MSE เมื่อ p เท่ากับ 0, 0.05, 0.10, 0.20, 0.30, 0.40

4. สรุปผลการทดลองและข้อเสนอแนะ

จากการศึกษาวิธีการเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนลของค่าเฉลี่ยถ่วงน้ำหนักและค่าเฉลี่ยตัวอย่าง โดยใช้ค่า MSE เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ วิธีที่มี MSE ต่ำสุด จะมีประสิทธิภาพดีที่สุด ผลการวิจัยสรุปได้ดังนี้ เมื่อข้อมูลมีค่าผิดปกติ โดย p เท่ากับ 0 หรือข้อมูลไม่มีค่าผิดปกติ $\bar{x}$ ยังคงเป็นวิธีที่มีประสิทธิภาพดีที่สุด ข้อมูลมีค่าผิดปกติในสัดส่วนที่น้อยคือ p เท่ากับ 0.05, 0.10 ค่าเฉลี่ยแบบถ่วงน้ำหนักด้วย วิธี plug-in เป็นวิธีที่มีประสิทธิภาพดีที่สุด และข้อมูลมีค่าผิดปกติในสัดส่วนที่มากคือ p เท่ากับ 0.20, 0.30, 0.40 ค่าเฉลี่ยแบบถ่วงน้ำหนักด้วยวิธี LSCV มีประสิทธิภาพดีที่สุด ดังนั้นตัวประมาณค่าเฉลี่ยแบบถ่วงน้ำหนักจึงเป็นทางเลือกหนึ่งในการประมาณค่าเฉลี่ยของประชากร

หากข้อมูลที่นำมาวิเคราะห์มีค่าผิดปกติ นอกจากนี้สามารถใช้เกณฑ์อื่นๆในการพิจารณาเปรียบเทียบประสิทธิตัวประมาณค่าเฉลี่ยแบบถ่วงน้ำหนัก เช่น MISE ค่าความแปรปรวน เป็นต้น

เอกสารอ้างอิง (References)

[1] ภทรวรรณ แสงนวกิจ, 2553. การประมาณค่าเฉลี่ยด้วยตัวประมาณแบบอัตราส่วน เมื่อข้อมูลมีค่าผิดปกติ. วิทยานิพนธ์ปริญญามหาบัณฑิต, สาขาสถิติประยุกต์และเทคโนโลยีสารสนเทศ สถาบันบัณฑิตพัฒนบริหารศาสตร์. [Patarawan Sangnawakij, 2010. Ratio Estimators of the Population Mean for Data with Outliers. M.S. Thesis, Applied Statistics and Information Technology, National Institute of Development Administration. (in Thai)]

[2] Huber, P.J., 1964. Robust estimation of a location parameter, *Annals of Mathematical Statistics,* 35, 73-101.

[3] Hampel, F.R., 1968. Contribution to the theory of robust estimation. Ph.D. thesis, University of California.

[4] กนกกาญจน์ รัตนไพบูลย์, 2546. ประสิทธิภาพของตัวประมาณค่าเฉลี่ยจากการถ่วงน้ำหนัก ด้วยค่าประมาณความหนาแน่นแบบเคอร์เนล. วิทยานิพนธ์ปริญญามหาบัณฑิต, สาขาสถิติประยุกต์, มหาวิทยาลัยศิลปากร. [Kanokkarn Rattanaphiboon, 2003. The Efficiency of the Mean Estimator Using Weight Based on Kernel Density Estimate. M.S. Thesis, Applied Statistics, Silpakorn University. (in Thai)]

[5] ปิยะฉัตร ลีลาศิลปะศาสน์, 2550. การเลือกแบนวิดจ์สำหรับการประมาณความหนาแน่นแบบเคอร์เนล ของฟังก์ชันตัวแปรสุ่ม. วิทยานิพนธ์ปริญญามหาบัณฑิต, สาขาสถิติประยุกต์, สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ. [Piyachat Leelasilapasart, 2006. Bandwidth Selection for Kernel Density Estimation for Function of Random Variable. M.S. Thesis, Applied Statistics, King Mongkut's Institute of Technology North Bangkok. (in Thai)]

[6] Wolfgang, H., 1990. Smoothing Techniques with Implementation in S. New York : Spinger-Verleg.

[7] Ahmad, I.A. and Mugdadi, A.R., 2004. A bandwidth selection for kernel density estimation of functions of random variables. *Computational Statistics & Data Analysis*, 47, 49 – 62.

[8] Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9, 65-78.

[9] Bowman, A.W., 1984. An alternative method of cross-validation for smoothing of density estimates. *Biometrika*, 71, 353-360.

[10] Duin, R.P.W., 1976. On the choice of smoothing parameters of Parzen estimators of probability density functions. *IEEE Transactions on Computers*, C-25, 1175–1179.

[11] Woodroofe, M., 1970. On choosing a delta sequence. *The Annals of Mathematical Statistics*, 41, 1665–1671.

[12] Silverman, B.W., 1998. Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC.