The Journal of Applied Science **Vol. 16 No. 1: 60-73 [2017]**
วารสารวิทยาศาสตร์ประยุกต์ doi: 10.14416/j.appsci.2017.06.002
ISSN 1513-7805 Printed in Thailand

Research Article

การเปรียบเทียบวิธีการประมาณค่าสูญหายในการสำรวจด้วยตัวอย่าง

A comparison of the estimation methods for missing data in sample survey

พิมพ์ชนก เชาวณาพรรณ์* และวัชรีภรณ์ ไชยมงคล

Pimchanok Chaovanaphan and Watchareeporn Chaimongkol

วิชาเอกสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ บางกะปิ กรุงเทพฯ 10240

Statistics Major, Graduate School of Applied Statistics, National Institute of Department Administration, Bangkapi, Bangkok 10240

***E-mail:** mint1234@hotmail.com

Received: 19/03/2017; Accepted: 6/06/2017

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะห่างต่ำสุด และวิธีเอ็มไอ โดยการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล ประกอบด้วยตัวแปรอิสระ และตัวแปรตามอย่างละหนึ่งตัวแปรที่มีการแจกแจงแบบปกติ มีค่าเฉลี่ย 10 และความแปรปรวน 1 สัมประสิทธิ์สหสัมพันธ์ คือ 0.10, 0.30, 0.50, 0.70 และ 0.90 กำหนดขนาดตัวอย่างเป็น 30, 60, 100 และ 300 ทำซ้ำ 1,000 รอบทุกกรณี กำหนดเปอร์เซ็นต์การสูญหายของตัวแปรตามเป็นไปอย่างสุ่มที่ 5, 10 และ 15 เปอร์เซ็นต์ ในทุกขนาดตัวอย่าง กำหนดสัมประสิทธิ์ความเชื่อมั่น 0.95 เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ สัมประสิทธิ์การแปรผันและค่าความน่าจะเป็นครอบคลุม ผลการศึกษา พบว่า เมื่อเปรียบเทียบค่าสัมประสิทธิ์การแปรผันของตัวประมาณ วิธีการประมาณค่า สูญหายทั้ง 3 วิธีให้ค่าสัมประสิทธิ์การแปรผันและค่าความน่าจะเป็นครอบคลุมของตัวประมาณใกล้เคียงกัน และค่าสัมประสิทธิ์การแปรผันของตัวประมาณมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น ส่วนค่าความน่าจะเป็นครอบคลุมของตัวประมาณมีค่าใกล้เคียงค่าสัมประสิทธิ์ความเชื่อมั่น 0.95 เนื่องจากวิธีการถดถอยเป็นวิธีที่ง่ายและไม่ซับซ้อน วิธีการถดถอยจึงเป็นวิธีการประมาณค่าสูญหายที่เหมาะสมที่สุดในการวิจัยครั้งนี้

คำสำคัญ: การประมาณค่า, ค่าสูญหาย

Abstract

The objective of this research was to study three imputation methods including, regression imputation method (RI), distance regression imputation method (DRI) and multiple imputation method (MI). The Monte

Carlo simulation technique, conducting for 1,000 replications, was composed of one independent variable (X) and one dependent variable (Y) with normal distribution, the mean of 10 and the variance of 1. The correlation coefficients were 0.10, 0.30, 0.50, 0.70 and 0.90. The sample sizes were 30, 60, 100 and 300. The Percentages of missing at random in the dependent variable were 5, 10 and 15. Confidence coefficient was 0.95. Coefficient of variation (CV) and coverage probability (CP) were used as the criteria of comparison. The result of this research showed that the three imputation methods led to similar coefficient of variations and coverage probabilities. When the sample sizes increased, the coefficient of variation decreased, and the coverage probability was adjacent to the confidence coefficient of 0.95. Since the Regression method was uncomplicated and easier, it was the most appropriate imputation method for this research.

Keywords: imputation, missing value

บทนำ

ปัจจุบันการเก็บรวบรวมข้อมูลของประชากรหรือการสำรวจด้วยตัวอย่าง มักจะพบปัญหาจากการไม่ตอบหรือไม่ให้ความร่วมมือของประชากรหรือกลุ่มตัวอย่าง ซึ่งทำให้ข้อมูลที่ได้จากประชากรหรือตัวอย่างสูญหาย โดยข้อมูลที่สูญหายนั้นอาจเป็นการไม่ตอบของข้อมูลบางหน่วย (unit nonresponse) หรือการไม่ตอบเฉพาะบางคำถาม (item nonresponse) จึงแก้ปัญหาเหล่านั้นด้วยการประมาณค่าข้อมูลที่เก็บไม่ได้ หรือการประมาณค่าข้อมูลสูญหาย (imputation) ให้สมบูรณ์ก่อน ถึงจะนำข้อมูลนั้นไปวิเคราะห์และสรุปผลได้ แต่มีนักวิจัยหรือผู้ทำการสำรวจบางกลุ่มที่แก้ปัญหาโดยการตัดหน่วยตัวอย่างที่สูญหายออกไป เหลือแต่หน่วยที่ครบถ้วน ซึ่งเป็นการกระทำที่ไม่เหมาะสม เนื่องจากจะก่อให้เกิดผลต่อการวิเคราะห์ข้อมูลหลายประการ ได้แก่ (1) เมื่อขนาดตัวอย่างลดลง ทำให้สูญเสียอำนาจในการทดสอบทางสถิติ (2) ค่าประมาณที่ได้อาจเป็นค่าที่เอนเอียง (3) รูปแบบของการสุ่มตัวอย่างเปลี่ยนไปทำให้ผลวิเคราะห์ที่ได้อาจไร้ความหมาย และ (4) การใช้น้ำหนักในการประมาณพารามิเตอร์อาจไม่มีความหมาย เนื่องจากผลบวกของน้ำหนักของหน่วยที่เหลืออยู่ขาดสมบัติที่ต้องการ หากไม่มีการปรับแก้ ดังนั้นการประมาณค่าของข้อมูลที่เก็บไม่ได้จึงมีความสำคัญมาก (Suwattee, 2009a; Suwattee, 2009b).

วิธีการประมาณค่าข้อมูลสูญหาย (imputation) มีมากมายหลายวิธี มีทั้งวิธีที่อาศัยตัวแบบในการประมาณค่าสูญหาย นั่นคือ model-donor imputation ซึ่งได้แก่ วิธีค่าเฉลี่ย (mean imputation) ที่เสนอโดย Wilks (1932) ซึ่งวิธีนี้เป็นการแทนค่าสูญหายด้วยค่าเฉลี่ยของตัวแปรที่สูญหายเมื่อตัดค่าสูญหายออกโดยไม่ใช้ตัวแปรช่วย ต่อมา Buck (1960) ได้นำเงื่อนไขของวิธีค่าเฉลี่ยมาใช้ในรูปแบบของวิธีการถดถอย (regression imputation) โดยการนำข้อมูลของตัวแปรที่สนใจกับตัวแปรช่วยมาใช้ในการประมาณค่าสูญหาย และอีกรูปแบบหนึ่งที่ใช้ข้อมูลของตัวแปรที่สนใจกับตัวแปรช่วย คือวิธีอัตราส่วน (ratio imputation) ซึ่งทั้งวิธีค่าเฉลี่ย วิธีสมการถดถอย และวิธีอัตราส่วนนั้น

เป็นวิธี single imputation คือ การแทนค่าสูญหายด้วยค่าเดียว ในปี 1987 Rubin ได้เสนอวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า (multiple imputation) เพื่อให้ได้ข้อมูลที่มีความถูกต้องสูง เนื่องจากมีการคำนวณหลายครั้งเพื่อให้ได้ค่าหลายค่า และนำค่ามาสรุปเพื่อให้ได้ค่าที่ดีที่สุด ลดข้อเสียของวิธี single imputation เนื่องจากข้อมูลที่ได้จากวิธี single imputation นั้นมีความแตกต่างจากค่าจริงและมีความเอนเอียงสูง (Rubin, 1987) และนอกจากนี้ยังมีวิธีการประมาณค่าสูญหายโดยไม่อาศัยตัวแบบ เรียกว่า real-donor imputation ซึ่งเป็นการประมาณค่าสูญหายที่ได้จากเซตข้อมูลของค่าที่สังเกตได้ ได้แก่ วิธี hot deck imputation, cold deck imputation และ nearest neighbor imputation เป็นต้น

ในงานวิจัยนี้ผู้วิจัยได้ทำการศึกษาการประมาณค่าสูญหายเฉพาะกรณีการไม่ตอบเฉพาะบางคำถาม (item nonresponse) และทำการประมาณค่าสูญหายด้วยวิธีการถดถอย (regression imputation : RI) วิธีการถดถอยด้วยระยะห่างต่ำสุด (distance regression imputation : DRI) และวิธีเอ็มไอ (multiple imputation: MI)

วิธีดำเนินการวิจัย

งานวิจัยนี้ใช้ข้อมูลจากการจำลองโดยวิธีมอนติคาร์โล (Monte carlo simulation) ด้วยโปรแกรม SAS สร้างประชากรขนาด 5,000 หน่วย ประกอบด้วยตัวแปรอิสระ (X) และตัวแปรตาม (Y) ที่มีการแจกแจงแบบปกติ มีค่าเฉลี่ย 10 และความแปรปรวน 1 มีระดับสหสัมพันธ์ (ρ) ระหว่าง X และ Y คือ 0.10, 0.30, 0.50, 0.70 และ 0.90 กำหนดขนาดตัวอย่างเท่ากับ 30, 60 100 และ 300 โดยสุ่มตัวอย่างแบบง่าย (simple random sampling) ทำซ้ำ 1,000 รอบในทุกกรณี กำหนดเปอร์เซ็นต์การสูญหายของตัวแปรตามให้สูญหายเป็นไปอย่างสุ่ม (missing at random) ที่ 5% 10% และ 15% ในทุกขนาดตัวอย่าง และกำหนดสัมประสิทธิ์ความเชื่อมั่น 0.95 เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวประมาณค่าเฉลี่ย คือ สัมประสิทธิ์การแปรผัน (coefficient of variation : CV) และค่าความน่าจะเป็นครอบคลุม (coverage probability)

การประมาณค่าสูญหายในการสำรวจด้วยตัวอย่างในครั้งนี้ ใช้วิธีการประมาณค่าข้อมูลสูญหาย 3 วิธี ดังต่อไปนี้

1. วิธีการถดถอย (regression imputation : RI)

การประมาณค่าสูญหายด้วยวิธีสมการถดถอย เป็นการประมาณค่าตัวแปรที่ต้องการศึกษาโดยใช้ความสัมพันธ์ระหว่างตัวแปรอิสระ (X) กับตัวแปรตาม (Y) มาช่วยในการประมาณค่า โดยกำหนดให้ $y_1,...,y_r$ เป็นค่าข้อมูลที่สมบูรณ์ และ $y_{r+1},...,y_n$ เป็นค่าข้อมูลที่สูญหาย และค่า x_i ทั้งหมดมีค่าที่สมบูรณ์ เมื่อ i = 1, .., n จากการใช้ข้อมูล $(x_1,y_1),...,(x_r,y_r)$ เพื่อประมาณค่าพารามิเตอร์ในตัวแบบ ดังนั้น จึงได้สมการถดถอยเพื่อประมาณค่าสูญหายเป็นดังนี้ (Jitthavech, 2015)

$$\hat{y}_i = b_o + b_1 x_i \quad (1)$$

 doi: 10.14416/j.appsci.2017.06.002

เมื่อ i = r+1, r+2, …, n

โดยที่ $\hat{y}_i$ เป็นค่าประมาณค่าสูญหายของตัวแปรตาม หน่วยที่ i

x_i เป็นค่าสังเกตของตัวแปรอิสระที่สอดคล้องกับตัวแปรตามหน่วยที่ i

b_0 , b_1 เป็นค่าประมาณของ β_0 และ β_1

เมื่อ $b_1 = \dfrac{\sum_{i=1}^{r}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{r}(x_i-\bar{x})^2}$, $b_0 = \bar{y} - b_1\bar{x}$, $\bar{x} = \dfrac{\sum_{i=1}^{r} x_i}{r}$ และ $\bar{y} = \dfrac{\sum_{i=1}^{r} y_i}{r}$

2. วิธีการถดถอยด้วยระยะทางต่ำที่สุด (distance regression imputation : DRI)

Chaimongkol (2005) ได้เสนอวิธีการประมาณค่าสูญหายด้วยวิธีการถดถอยระยะห่างต่ำสุด หรือ DRI โดยได้กำหนดให้ $y_1,...,y_r$ เป็นค่าข้อมูลที่สมบูรณ์ และ $y_{r+1},...,y_n$ เป็นค่าข้อมูลที่สูญหาย และค่า x_i i = 1, .., n ทั้งหมดมีค่า เมื่อ $i = 1,...,r$ ให้ $d_{xi} = x_{(i+1)} - x_{(i)}$ และ $d_{yi} = y'_{i+1} - y'_i$ เมื่อ y'_i คือค่าของ y ที่สอดคล้องกับสถิติสำดับของ $x_{(i)}$ ดังนั้น สมการถดถอยเชิงเส้นอย่างง่ายระหว่าง d_{xi} กับ d_{yi} คือ

$$\hat{d}_y = \hat{\beta}'_o + \hat{\beta}'_1 d_x \tag{2}$$

เมื่อ $\hat{\beta}'_1 = \dfrac{\sum_{i=1}^{r-1}(d_{xi}-\bar{d}_x)(d_{yi}-\bar{d}_y)}{\sum_{i=1}^{r-1}(d_{xi}-\bar{d}_x)^2}$, $\hat{\beta}'_o = \bar{d}_y - \hat{\beta}_1'\bar{d}_x$, $\bar{d}_y = \dfrac{1}{r-1}\sum_{i=1}^{r-1} d_{yi}$ และ $\bar{d}_x = \dfrac{1}{r-1}\sum_{i=1}^{r-1} d_{xi}$

สำหรับหน่วยที่มีค่าสูญหาย j = r+1, …, n

ให้ $m_j = |x_j - x_{(i)}| = \min_{1\le k\le r} |x_i - x_{(k)}|$ สำหรับบางค่า k เมื่อ $1 \le k \le r$ ดังนั้น

$$\Delta y_j = \hat{\beta}'_0 + \hat{\beta}'_1 m_j \tag{3}$$

เมื่อ Δy_j เป็นค่าประมาณระยะห่างของ y_j ดังนั้น ค่าสูญหายจะถูกแทนที่ด้วยค่าประมาณ $\hat{y}'_j = y'_k + \Delta y_i$ เมื่อ y'_k เป็นค่าของ y ที่สอดคล้องกับค่า $x_{(k)}$ โดยสถิติลำดับที่ k มีค่าใกล้เคียงกับค่า $x_{(i)}$ เมื่อ $1 \le k \le r$

โครงสร้างของวิธี DRI

 doi: 10.14416/j.appsci.2017.06.002

$x_{(i)}$	d_{xi}	y'_i	d_{yi}	Δy_j	$\hat{y}'_i$
$x_{(1)}$	$x_{(2)} - x_{(1)}$	y'_1	$y'_2 - y'_1$		y'_1
$x_{(2)}$	$x_{(3)} - x_{(2)}$	y'_2	$y'_3 - y'_2$		y'_2
…	…	…	…		…
$x_{(r-1)}$	$x_{(r)} - x_{(r-1)}$	y'_{r-1}	$y'_r - y'_{r-1}$		y'_{r-1}
$x_{(r)}$	-	y'_r	-		y'_r
x_{r+1}				Δy_{r+1}	$\hat{y}'_k + \Delta y_{r+1}$
…				…	…
x_n				Δy_n	$\hat{y}'_k + \Delta y_n$

ตัวประมาณค่าสูญหายของ y_j เมื่อ $j = r+1, \ldots, n$ แสดงได้ดังนี้

สำหรับ $\hat{\beta}'_1 > 0$

$$\hat{y}'_j = \begin{cases} y'_k + \Delta y_j \text{ เมื่อ } x_j \geq x_{(k)} \\ y'_k - \Delta y_j \text{ เมื่อ } x_j < x_{(k)} \end{cases} \quad (4)$$

และสำหรับ $\hat{\beta}'_1 < 0$

$$\hat{y}'_j = \begin{cases} y'_k + \Delta y_j \text{ เมื่อ } x_j \leq x_{(k)} \\ y'_k - \Delta y_j \text{ เมื่อ } x_j > x_{(k)} \end{cases} \quad (5)$$

3. วิธีเอ็มไอ (multiple imputation)

วิธีการประมาณค่าสูญหายด้วยวิธีเอ็มไอนั้น ได้มีการเสนอแนวความคิดโดย Rubin (1987) เป็นการแทนที่ค่าสูญหายด้วยวิธีการต่าง ๆ ตั้งแต่ 2 วิธีขึ้นไป (รูปที่ 1) มีขั้นตอนดังนี้ (Little & Rubin, 1987, Carpenter & Kenward, 2013)

3.1 สร้างชุดข้อมูลที่สมบูรณ์ด้วยวิธีการประมาณค่าสูญหายแต่ละวิธี

3.2 วิเคราะห์ตัวประมาณค่าพารามิเตอร์ที่สนใจศึกษาจากแต่ละวิธีการประมาณค่าสูญหาย

3.3 รวมตัวประมาณค่าพารามิเตอร์จากแต่ละวิธีเป็นชุดเดียวกัน โดยใช้กฎของ Rubin นั่นคือค่าประมาณของตัวประมาณค่าเฉลี่ย คือ

doi: 10.14416/j.appsci.2017.06.002

$$\overline{\theta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i \tag{6}$$

เมื่อ $\hat{\theta}_i$ แทน ค่าเฉลี่ยของข้อมูลของแต่ละวิธี ; $i = 1, 2, ..., M$

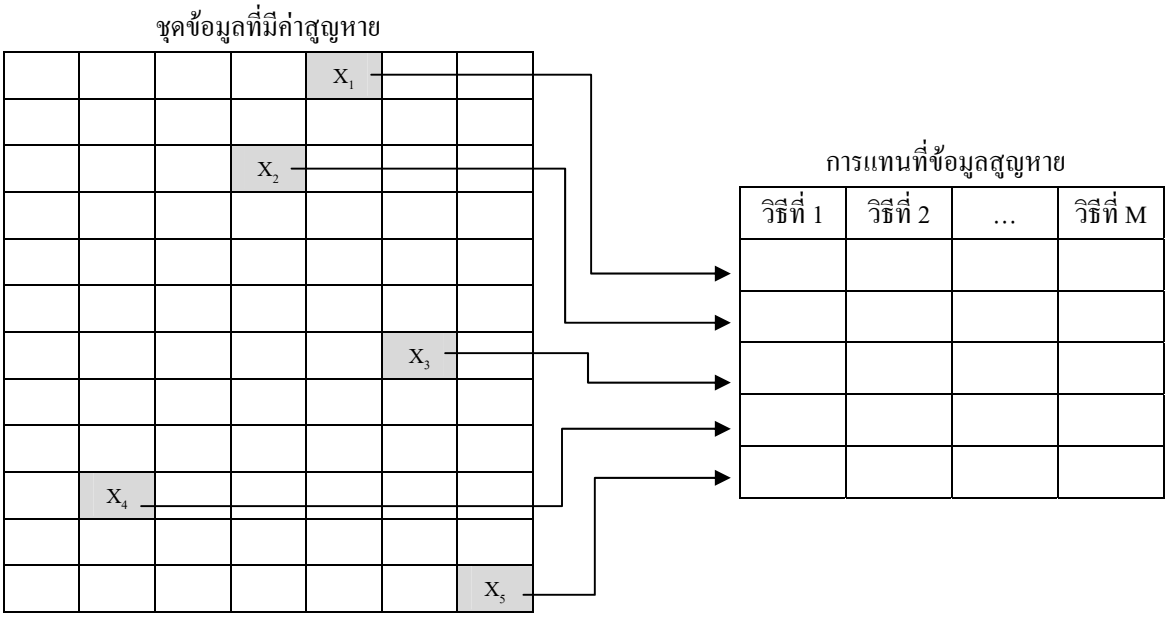
ความแปรปรวนของตัวประมาณค่าเฉลี่ย เกิดจากองค์ประกอบ 2 องค์ประกอบ คือความแปรปรวนภายใน และความแปรปรวนภายนอก

ความแปรปรวนภายใน $$\overline{W} = \frac{1}{M}\sum_{i=1}^{M} W_i \tag{7}$$

เมื่อ W_i แทน ความแปรปรวนของแต่ละวิธี ; $i = 1, 2, ..., M$

ความแปรปรวนภายนอก $$B = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\theta}_i - \overline{\theta})^2 \tag{8}$$

ดังนั้น ความแปรปรวนของตัวประมาณค่าเฉลี่ย คือ $T = \overline{W} + (1 + M^{-1})B$



รูปที่ 1. วิธีการแทนที่ข้อมูลสูญหายด้วยวิธีเอ็มไอ

 doi: 10.14416/j.appsci.2017.06.002

เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของตัวประมาณ 2 วิธี คือ

1. สัมประสิทธิ์การแปรผัน (coefficient of variation)

$$c.v.(\hat{\theta}) = \frac{\sqrt{v(\hat{\theta})}}{\mu_\theta} \times 100 \quad (9)$$

2. ค่าความน่าจะเป็นครอบคลุม (coverage probability)

เมื่อช่วงความเชื่อมั่น 95% ของ $\bar{y}_I$ สร้างจาก $\bar{y}_I \pm 1.96\sqrt{\hat{V}(\bar{y}_I)}$ และค่าความน่าจะเป็นครอบคลุม คือ

$$CP_M(\bar{y}_I) = \frac{t}{M} \quad (10)$$

โดยที่ $\bar{y}_I$ คือ ค่าเฉลี่ยของ y เมื่อมีการทดแทนค่าสูญหายแล้ว

M คือ จำนวนรอบในการทำซ้ำ ; M = 1,000 รอบ

t คือ จำนวนครั้งที่ช่วงความเชื่อมั่นคลุมค่าเฉลี่ยของ Y

การทดสอบของตูกีย์ (Tukey'sTest)

การทดสอบของตูกีย์ (Tukey's test) เสนอโดย Tukey เป็นการทดสอบความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มหรือเปรียบเทียบค่าเฉลี่ยแต่ละคู่โดยการคำนวณค่าวิกฤตในการเปรียบเทียบค่าเฉลี่ยทุกคู่ที่เป็นไปได้นั้น จะใช้ค่านัยสำคัญของพิสัยสติวเด็นท์ไตซ์ (significant studentized range) นั่นคือ

$$T_\alpha = q_\alpha(k, \upsilon)s_{\bar{y}} \quad 11)$$

เมื่อ $q_\alpha(k, \upsilon)s_{\bar{y}}$ คือ ค่านัยสำคัญของพิสัยสติวเด็นท์ไตซ์ที่ระดับนัยสำคัญ α และที่องศาความเสรีของความคลาดเคลื่อนเท่ากับ υ ในการวิเคราะห์ k กลุ่ม

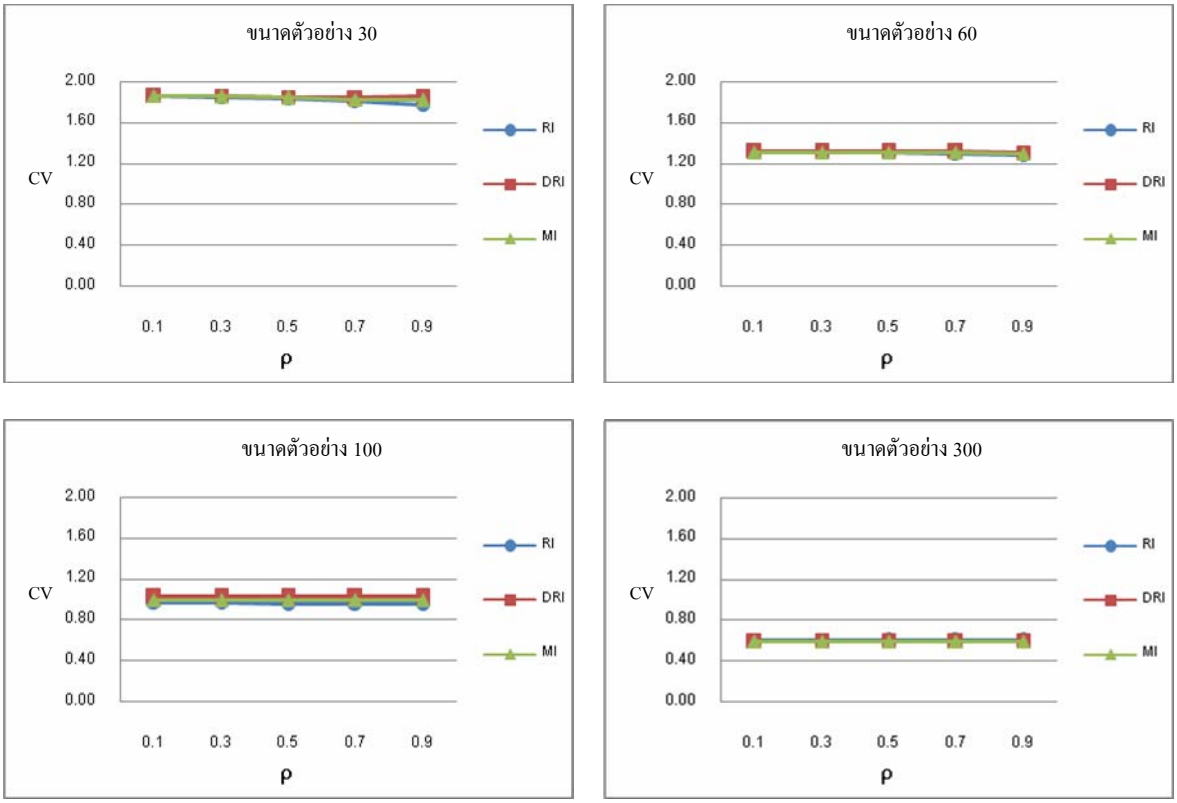
doi: 10.14416/j.appsci.2017.06.002

ผลการวิจัย

1. ผลการวิเคราะห์การเปรียบเทียบสัมประสิทธิ์การแปรผันของตัวประมาณ

1.1 ผลการวิเคราะห์การเปรียบเทียบสัมประสิทธิ์การแปรผันของตัวประมาณ เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 5%

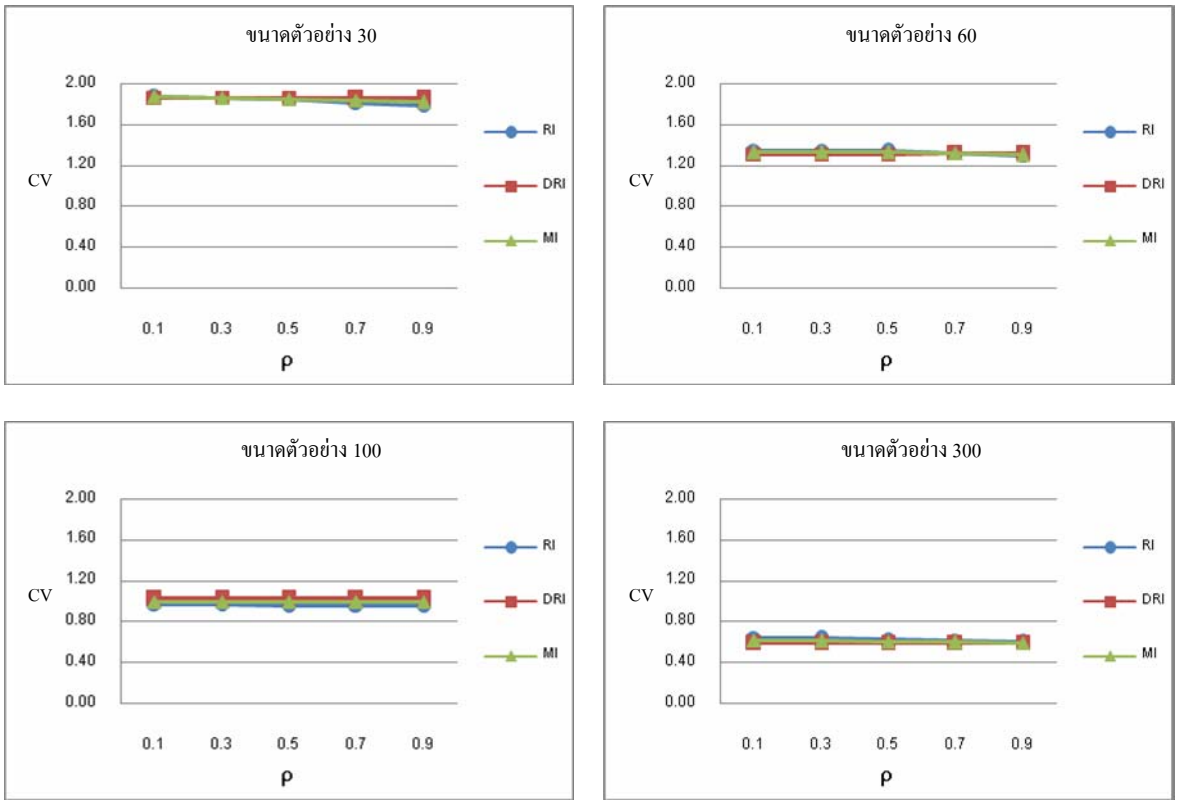
พิจารณาค่าสัมประสิทธิ์การแปรผันของตัวประมาณ ที่เปอร์เซ็นต์การสูญหายของข้อมูล 5% พบว่า เมื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีเอ็มไอ ด้วยการทดสอบของตูกีย์ ให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณใกล้เคียงกัน ไม่ว่าขนาดตัวอย่างและสัมประสิทธิ์สหสัมพันธ์จะเป็นเท่าใดก็ตาม และเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่าสัมประสิทธิ์การแปรผันของตัวประมาณมีแนวโน้มลดลงในทุกกรณี (รูปที่ 2)



รูปที่ 2. ค่าสัมประสิทธิ์การแปรผันของตัวประมาณ เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 5% จำแนกตามขนาดตัวอย่าง

1.2 ***ผลการวิเคราะห์การเปรียบเทียบสัมประสิทธิ์การแปรผันของตัวประมาณ เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 10%***

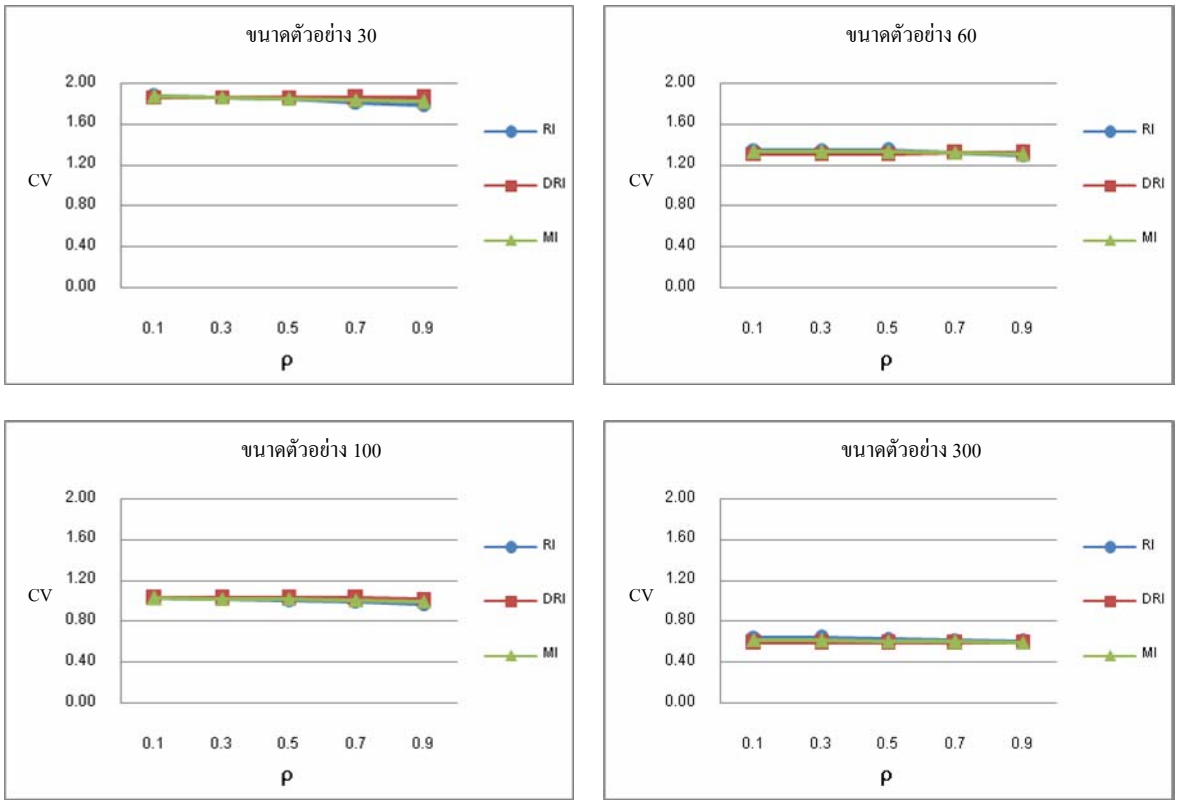
พิจารณาค่าสัมประสิทธิ์การแปรผันของตัวประมาณ ที่เปอร์เซ็นต์การสูญหายของข้อมูล 10% พบว่า เมื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีด้วยการทดสอบของตูกีย์ ให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณใกล้เคียงกัน (รูปที่ 3)



รูปที่ 3. ค่าสัมประสิทธิ์การแปรผันของตัวประมาณ เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 10% จำแนกตามขนาดตัวอย่าง

1.3 ***ผลการวิเคราะห์การเปรียบเทียบสัมประสิทธิ์การแปรผันของตัวประมาณ เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 15%***

พิจารณาค่าสัมประสิทธิ์การแปรผันของตัวประมาณ ที่เปอร์เซ็นต์การสูญหายของข้อมูล 15% พบว่า เมื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายด้วยวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีเอ็มไอ ด้วยการทดสอบของตูกีย์ ให้ค่าสัมประสิทธิ์การแปรผันของตัวประมาณใกล้เคียงกัน และค่าสัมประสิทธิ์การแปรผันของตัวประมาณมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น (รูปที่ 4)

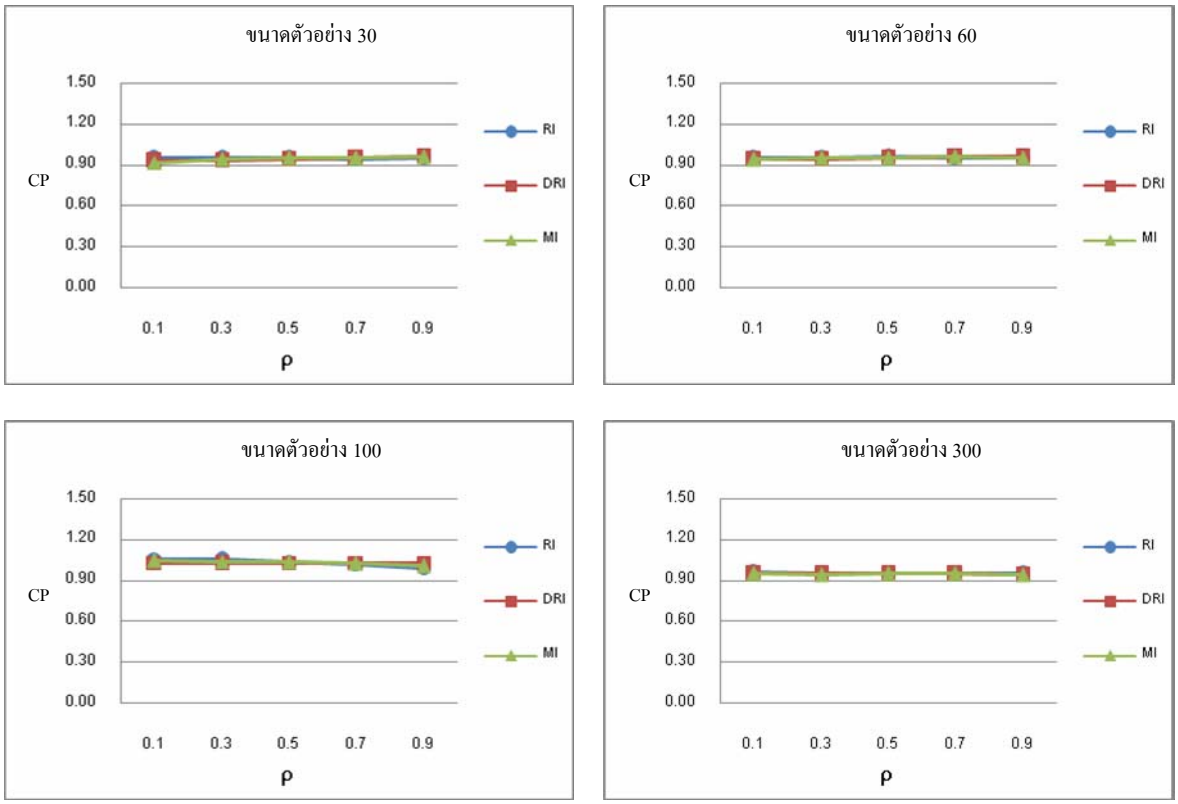


รูปที่ 4. ค่าสัมประสิทธิ์การแปรผันของตัวประมาณ เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 15% จำแนกตามขนาดตัวอย่าง

2. แสดงผลการวิเคราะห์การเปรียบเทียบค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร

2.1 ผลการวิเคราะห์การเปรียบเทียบค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 5%

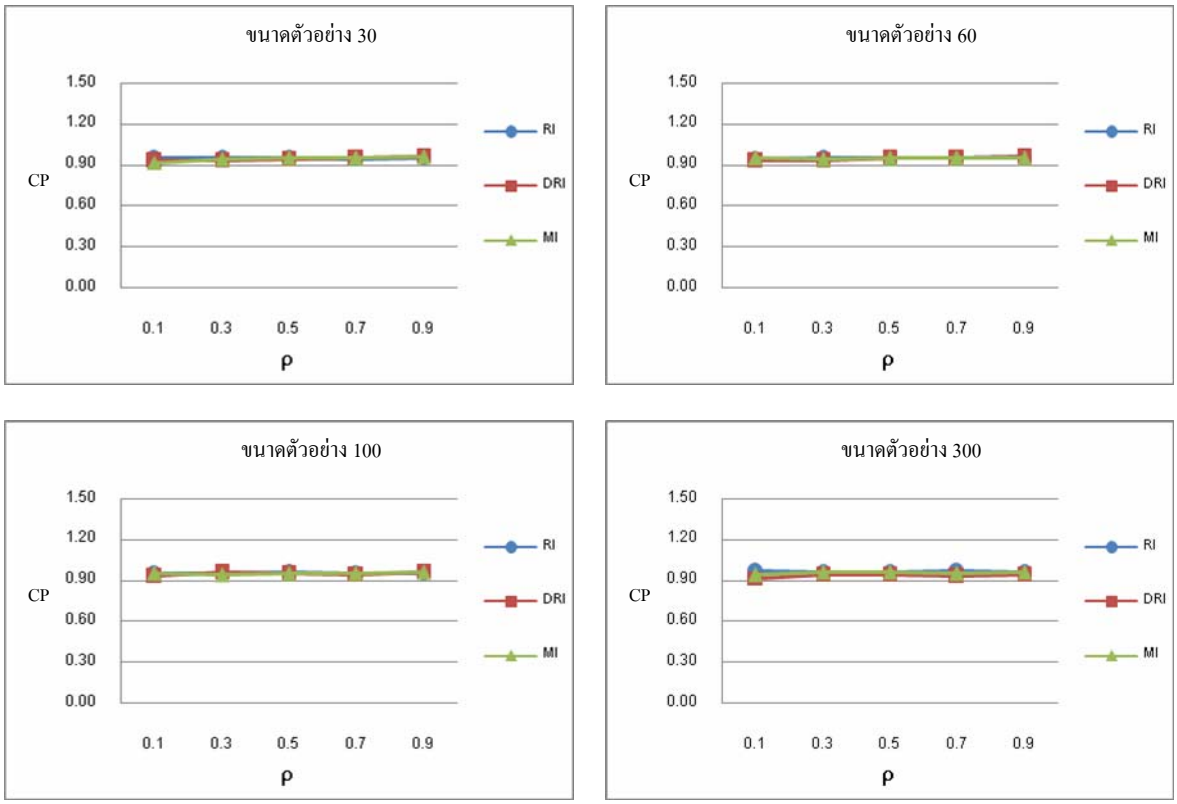
พิจารณาค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร ที่เปอร์เซ็นต์การสูญหายของข้อมูล 5% พบว่า เมื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีเอ็มไอ ด้วยการทดสอบของตูกีย์ ให้ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากรไม่แตกต่างกัน มีค่าใกล้เคียงค่าสัมประสิทธิ์ความเชื่อมั่น 0.95 (รูปที่ 5)



รูปที่ 5. ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 5% จำแนกตามขนาดตัวอย่าง

2.2 ผลการวิเคราะห์การเปรียบเทียบค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 10%

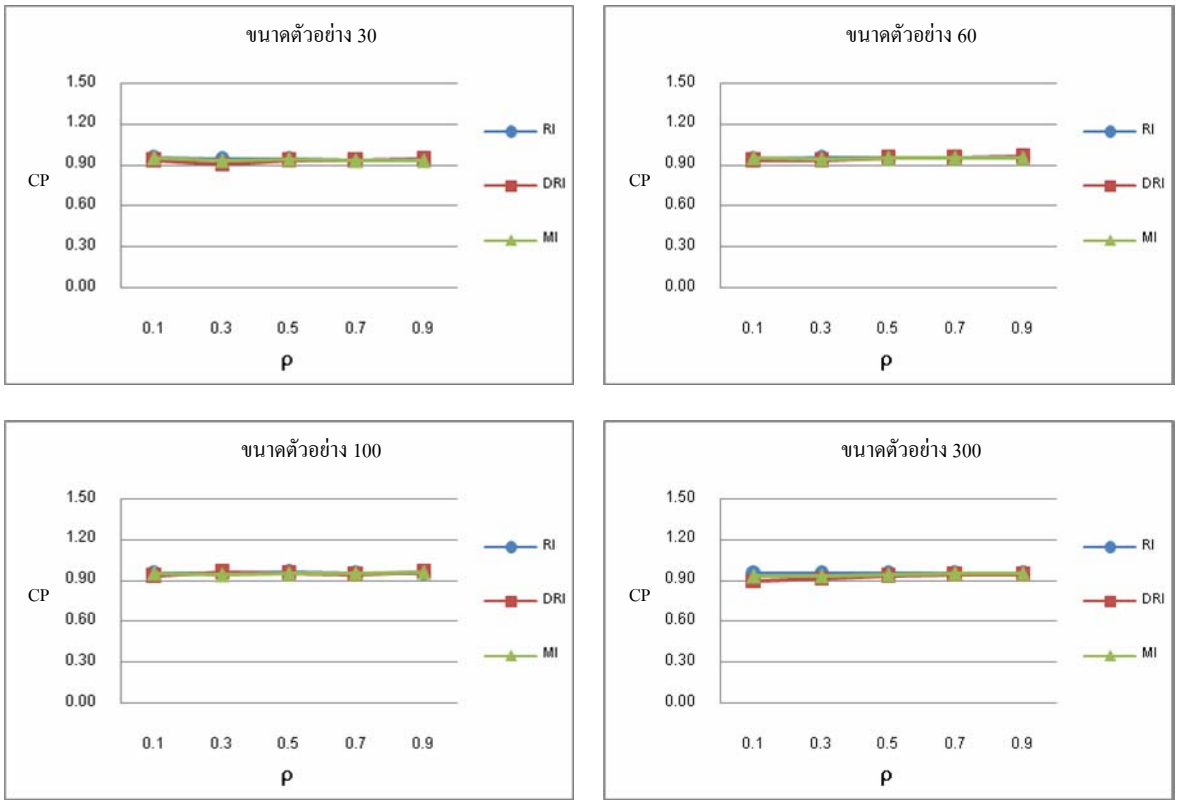
พิจารณาค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร ที่เปอร์เซ็นต์การสูญหายของข้อมูล 10% พบว่า เมื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีเอ็มไอ ด้วยการทดสอบของตูกีย์ ให้ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากรไม่แตกต่างกัน มีค่าใกล้เคียงค่าสัมประสิทธิ์ความเชื่อมั่น 0.95 (รูปที่ 6)



รูปที่ 6. ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 10% จำแนกตามขนาดตัวอย่าง

2.3 ผลการวิเคราะห์การเปรียบเทียบค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 15%

พิจารณาค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร ที่เปอร์เซ็นต์การสูญหายของข้อมูล 15% พบว่า เมื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีเอ็มไอ ด้วยการทดสอบของตูกีย์ ให้ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากรไม่แตกต่างกัน มีค่าใกล้เคียงค่าสัมประสิทธิ์ความเชื่อมั่น 0.95 (รูปที่ 7)



รูปที่ 7. ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากร เมื่อเปอร์เซ็นต์การสูญหายของข้อมูล 15% จำแนกตามขนาดตัวอย่าง

วิจารณ์และสรุปผลการวิจัย

วิธีการประมาณค่าสูญหายในการสำรวจด้วยตัวอย่างนั้น มีหลายวิธีด้วยกัน ในงานวิจัยนี้ ศึกษาวิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีเอ็มไอ ผลการศึกษา พบว่า วิธีการประมาณค่าสูญหายทั้ง 3 วิธี ให้ค่าสัมประสิทธิ์การแปรผันและค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากรที่ใกล้เคียงกันหรือไม่แตกต่างกัน โดยที่ค่าสัมประสิทธิ์การแปรผันของตัวประมาณมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น เนื่องจากข้อมูลที่มีขนาดใหญ่จะช่วยลดความคลาดเคลื่อนหรือเพิ่มความแม่นยำในการประมาณมากขึ้น ขณะที่ค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากรมีค่าใกล้เคียงค่าสัมประสิทธิ์ความเชื่อมั่น 0.95 และวิธีการถดถอยเป็นวิธีที่ง่ายและไม่ซับซ้อน วิธีการถดถอยจึงเป็นวิธีการประมาณค่าสูญหายที่เหมาะสมที่สุดในการวิจัยครั้งนี้

เมื่อพิจารณาขนาดตัวอย่าง เปอร์เซ็นต์การสูญหายของข้อมูล และค่าสัมประสิทธิ์สหสัมพันธ์ พบว่า ระดับของค่าสัมประสิทธิ์สหสัมพันธ์ ไม่มีผลต่อการเลือกวิธีการประมาณค่าสูญหาย เนื่องจากในแต่ละกรณีเมื่อค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเปลี่ยนแปลงไปค่าสัมประสิทธิ์การแปรผันและค่าความน่าจะเป็นครอบคลุมค่าเฉลี่ยของประชากรมีค่าใกล้เคียงกัน ขณะที่ขนาดตัวอย่างเพิ่มขึ้นค่าสัมประสิทธิ์การแปรผันมีค่าลดลง ผลสรุปที่ได้จากการเปรียบเทียบค่าสัมประสิทธิ์การแปรผันแตกต่างจากงานวิจัยของ Chaimongkol (2005) เป็นผลมาจากงานวิจัยของ Chaimongkol (2005) ศึกษาข้อมูลที่มีการแจกแจงแบบยูนิฟอร์มและการแจกแจงแบบแกมมา ขณะที่งานวิจัยนี้ใช้การแจกแจงแบบปกติ

เอกสารอ้างอิง

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological), 22*(2): 302-306.

Carpenter, J. R. & Kenward, M. G. (2013). *Multiple Imputation and Its Application* (1st ed.). West Sussex, UK: John Wiley & Sons, Ltd.

Chaimongkol, W. (2005). *Three composite imputation methods for item nonresponse estimation in sample surveys* (Doctoral dissertation) Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok.

Jitthavech J. (2015). *Regression Analysis* (1st ed.). Bangkok, Thailand: Academic Promotion and Development Program, National Institute of Development Administration. (in Thai)

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data* (1st ed.). NewYork: John Wiley & Sons, Ltd.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys* (1st ed.). New York: John Wiley & Sons, Ltd.

Suwattee P. (2009a). *Sample Surveys : Sampling Designs and Analysis* (1st ed.). Bangkok, Thailand: Academic Promotion and Development Program, National Institute of Development Administration. (in Thai)

Suwattee P. (2009b). *Sampling Theory* (1st ed.). Bangkok, Thailand: Academic Promotion and Development Program, National Institute of Development Administration. (in Thai)

Wilks, S. S., (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, *3*(3): 163-195. doi:10.1214/aoms/1177732885