ว.วิทย. มข. 41(1) 239-249 (2556) KKU Sci. J. 41(1) 239-249 (2013)

สมการสำหรับการคำนวณค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่อง

Equation for Rapid Calculation of the Standard Deviation

พงษ์วุฒิ ดวงศรี[1]

บทคัดย่อ

สมการสำหรับการคำนวณค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่อง เป็นสมการเพื่อการคำนวณค่าเบี่ยงเบนมาตรฐานของข้อมูล ที่มีจำนวนค่าสังเกตเพิ่มขึ้นไม่แน่นอน โดยการคำนวณไม่จำเป็นต้องคำนวณจากค่าสังเกตทั้งหมดทุกครั้งเมื่อมีค่าสังเกตเพิ่มเข้ามาใหม่ แต่จะสามารถคำนวณหาค่าได้อย่างต่อเนื่องจากสมาชิกหนึ่งไปยังอีกสมาชิกหนึ่ง ดังนั้นจึงทำให้เกิดประสิทธิภาพการคำนวณในลักษณะ บิ๊กโอ-เอ็น O(N) แทนที่จะเป็นบิ๊กโอ-เอ็นกำลังสอง O(N^2) อย่างที่เคยคำนวณหาค่าในรูปแบบของสมการมาตรฐานทั่วไป

สำหรับสมการค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่องที่สร้างขึ้น รูปแบบของสมการจะขึ้นอยู่กับค่าสังเกตที่เพิ่มมาใหม่ (x_N) ขนาดข้อมูลปัจจุบัน (N) ค่าเฉลี่ยข้อมูลปัจจุบัน (μ) และค่าเบี่ยงเบนมาตรฐานของข้อมูลก่อนหน้านี้ (σ') รูปลักษณ์ของสมการอยู่ในรูปแบบกะทัดรัด ไม่ซับซ้อนเมื่อเทียบกับรูปแบบอื่น ง่ายต่อการเข้าใจและง่ายในการนำไปใช้

ABSTRACT

This paper proposes rapid standard deviation equation for determining the standard deviation of variable-size data. The novel method calculates the standard deviation in an accumulative way: it requires only new observation to be included, instead of using whole data, in calculation. The method has the running time of O(N) instead of the traditional O(N^2).

The proposed rapid standard deviation equation depends on the following factors: new observation (x_N), the amount of current data size (N), the average of the value of current data (μ), and standard deviation of previous data set (σ'). The equation is in a compact form, easy to understand and apply to data sets.

[1]คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ตำบลแสนสุข อำเภอเมือง จังหวัดชลบุรี 20131

E-mail: phong@buu.ac.th

คำสำคัญ: สมการสำหรับการคำนวณค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่อง

Keywords: Rapid standard deviation equation

คำนำ

ในการวัดการกระจายของข้อมูล โดยทั่วไปจะใช้ส่วนเบี่ยงเบนมาตรฐานในการศึกษา ซึ่งสมการการหาค่าเบี่ยงเบนมาตรฐานที่เป็นมาตรฐาน (กัลยา, 2553; Bartz, 1999) และใช้กันอย่างแพร่หลายคือ

สำหรับข้อมูลตัวอย่าง $$S_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad ; \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(I)}$$

สำหรับข้อมูลประชากร $$\sigma_N = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}} \quad ; \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \text{(II)}$$

โดยที่ x_i หมายถึง ค่าสังเกตหรือข้อมูลค่าที่ I โดย I = 1, 2, ... n

n หมายถึง จำนวนค่าสังเกต หรือขนาดข้อมูลของตัวอย่าง

N หมายถึง จำนวนค่าสังเกต หรือขนาดข้อมูลของประชากร

$\bar{x}$ หมายถึง ค่าเฉลี่ยของค่าสังเกตหรือข้อมูลตัวอย่าง

μ หมายถึง ค่าเฉลี่ยของค่าสังเกตหรือข้อมูลประชากร

จากสมการจะเห็นว่า การคำนวณหาค่าต่าง ๆ จะต้องทราบขนาดข้อมูลทั้งหมด และการหาค่าจะคำนวณจากค่าสังเกตของข้อมูลทั้งหมด ดังนั้นในบางกรณี หรือบางปัญหา ที่มีขนาดข้อมูลเปลี่ยนแปลง ไม่คงที่ และจำเป็นต้องหาค่าเบี่ยงเบนมาตรฐานปัจจุบัน เพื่อนำไปใช้ในครั้งต่อไป วิธีการนี้จึงไม่เหมาะสมและไม่มีประสิทธิภาพ เพราะต้องคำนวณจาก ค่าสังเกตของข้อมูล ใหม่ทั้งหมดทุกครั้ง ตัวอย่างเช่นการแบ่งกลุ่มข้อมูลในสาขาวิชาฟัซซีลอจิก (พยุง, 2553; สืบค้นจาก http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=801205) ที่อาศัยการหาระยะห่างระหว่างค่าที่เข้ามาใหม่กับจุดอ้างอิง (euclidean distance) และอาศัยการแจกแจงแบบปกติหรือการแจกแจงแบบเกาส์เซียน (กัลยา, 2553; Bruning and Kintz, 1997) (normal or Gaussian distribution) $f(x_i) = \exp\left(-\frac{\|x_i - \mu_i'\|^2}{2\sigma_i'^2}\right)$ เพื่อพิจารณาว่าค่าสังเกตที่เข้ามาใหม่ อยู่ภายใต้การแจกแจงนี้หรือไม่ ถ้าเป็นที่ยอมรับได้ ก็ให้รวมค่าสังเกตนี้เข้ากลุ่ม แล้วหาค่าเบี่ยงเบนมาตรฐานของกลุ่มค่าสังเกตนี้ใหม่เพื่อใช้ในการพิจารณาครั้งต่อไป แต่ถ้าไม่สามารถเข้ากลุ่มนี้ได้ ก็ไปทดลองหากลุ่มอื่น และถ้าไม่สามารถเข้ากลุ่มใดได้ ก็ให้แยกเป็นกลุ่มใหม่โดยมีค่าสังเกตนี้ เป็นค่าแรกของกลุ่ม กับค่าเบี่ยงเบนมาตรฐานที่สุ่มมาสำหรับกลุ่มใหม่นี้

ด้วยเหตุดังกล่าว จึงมีความต้องการในการสร้างสมการหาค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่อง เพื่อหลีกเลี่ยงการคำนวณซ้ำซ้อน และสามารถคำนวณได้อย่างต่อเนื่อง โดยยึดหลักการว่าสมการต้องอยู่ในรูปแบบไม่ซับซ้อน ง่ายในการเข้าใจ และสะดวกต่อการใช้งาน

การดำเนินงาน

เนื่องจากเป็นการสร้างสมการ ดังนั้นจึงไม่จำเป็นต้องใช้อุปกรณ์อะไร นอกเหนือจากคอมพิวเตอร์ส่วนบุคคลและชุดโปรแกรมสำนักงาน สำหรับขั้นตอนการดำเนินงานประกอบด้วย การสร้างสมการ การตรวจสอบความถูกต้อง และการวิเคราะห์ประสิทธิภาพของสมการ โดยอาศัยอัลกอริธึมการคำนวณหาค่าเบี่ยงเบนมาตรฐาน (สืบค้นจาก http://www.answers.com/topic/algorithms-for-calculating-variance) ซึ่งมีรายละเอียดดังนี้

กำหนดให้: x_i เป็นค่าสังเกตที่เป็นจำนวนจริง ลำดับที่ i $\{x_i \in \mathrm{R} \mid i = 1, 2, 3, ..., n', n\}$; $\{n \in \mathrm{I}^+ \mid \mathrm{I}^+ \geq 1\}$

n เป็นจำนวนเต็มบวกที่มีจำนวนจำกัด สำหรับในกรณีนี้เป็นขนาดข้อมูลทั้งหมด n ค่า

x_n เป็นค่าสังเกตลำดับที่ n หรือเป็นลำดับสุดท้ายของชุดข้อมูล

n' เป็นลำดับของค่าสังเกต ก่อนหน้าลำดับที่ n เช่น ข้อมูล $x_{n'}$ หรือขนาดข้อมูล n - 1 ค่า

$\therefore\ n = n' + 1$ หรือ $n' = n - 1$ (III)

การหาค่าเฉลี่ยของข้อมูลทั้งหมด (Navidi, 2011) n ค่า สามารถคำนวณได้จากสมการดังต่อไปนี้

การคำนวณค่าจากข้อมูลทั้งหมด: ค่าเฉลี่ยของข้อมูล n ค่า $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$; ค่าเฉลี่ยของข้อมูล n' ค่า $\overline{x'} = \dfrac{\sum_{i=1}^{n-1} x_i}{n-1}$

หรือ การคำนวณค่าแบบต่อเนื่อง: $\bar{x} = \dfrac{\overline{x'}n' + x_n}{n}$ หรือ $\overline{x'} = \dfrac{\bar{x}n - x_n}{n'}$ (IV)

1. สมการหาค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่องสำหรับข้อมูลกลุ่มตัวอย่าง

จากที่กำหนดให้ ถ้า x_i คือค่าสังเกตของกลุ่มตัวอย่างที่มีขนาดข้อมูลจำกัดทั้งหมด n ค่า การคำนวณหาค่าเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่าง สามารถหาได้จากสมการมาตรฐานทั่วไปดังนี้

ค่าเบี่ยงเบนมาตรฐาน $$S_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad [1]$$

ซึ่งสามารถเขียนได้ดังนี้ $$S_n = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_{n'} - \bar{x})^2 + (x_n - \bar{x})^2}{n-1}} \quad (1)$$

เมื่อพิจารณาเฉพาะส่วน $(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_{n'} - \bar{x})^2$ จะได้ดังนี้

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_{n'}-\bar{x})^2 =$$
$$(x_1^2-2x_1\bar{x}+\bar{x}^2)+(x_2^2-2x_2\bar{x}+\bar{x}^2)+\ldots+(x_{n'}^2-2x_{n'}\bar{x}+\bar{x}^2)$$
$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_{n'}-\bar{x})^2 =$$
$$(x_1^2-2x_1\bar{x'}+\bar{x'}^2)+(x_2^2-2x_2\bar{x'}+\bar{x'}^2)+\ldots+(x_{n'}^2-2x_{n'}\bar{x'}+\bar{x'}^2)+$$
$$\{(-2x_1\bar{x}+\bar{x}^2)+(-2x_2\bar{x}+\bar{x}^2)+\ldots+(-2x_{n'}\bar{x}+\bar{x}^2)\}-$$
$$\{(-2x_1\bar{x'}+\bar{x'}^2)+(-2x_2\bar{x'}+\bar{x'}^2)+\ldots+(-2x_{n'}\bar{x'}+\bar{x'}^2)\} \quad (2)$$

โดย $(x_1^2-2x_1\bar{x'}+\bar{x'}^2)+(x_2^2-2x_2\bar{x'}+\bar{x'}^2)+\ldots+(x_{n'}^2-2x_{n'}\bar{x'}+\bar{x'}^2) =$

$$\frac{(n'-1)\{(x_1-\bar{x'})^2+(x_2-\bar{x'})^2+\ldots+(x_{n'}-\bar{x'})^2\}}{(n'-1)} = (n'-1)s'^2 \quad (V)$$

จะเห็นว่า (V) จะเป็นจริงก็ต่อเมื่อ $n > 2$ เพราะถ้า $n = 2$ แล้ว n'-1 = 0 และที่ n'= 1. ค่า S'ก็ไม่มีค่าเช่นกัน

แทน (V) ลงใน (2)

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_{n'}-\bar{x})^2$$
$$= (n'-1)s'^2+\{n'\bar{x}^2-2\bar{x}(x_1+x_2+..x_{n'})\}-\{n'\bar{x'}^2-2\bar{x'}(x_1+x_2+..x_{n'})\}$$
$$= (n'-1)s'^2+\{n'\bar{x}^2-2\bar{x}.n'\bar{x'}\}-\{n'\bar{x'}^2-2\bar{x'}n'\bar{x'}\}$$
$$= (n'-1)s'^2+n'\bar{x}^2-2n'\bar{x}.\bar{x'}+n'\bar{x'}^2 \quad (3)$$

แทน (3) และ n'= n – 1 ลงใน (1)

จะได้ $$S_n=\sqrt{\frac{(n'-1).s'^2+n'\bar{x}^2-2n'\bar{x}.\bar{x'}+n'\bar{x'}^2+(x_n-\bar{x})^2}{n-1}}$$

$$S_n=\sqrt{\frac{(n'-1).s'^2+n'(\bar{x}-\bar{x'})^2+(x_n-\bar{x})^2}{n'}}$$

หรือ $$S_n=\sqrt{(1-\frac{1}{n'})s'^2+\frac{n'\bar{x}^2-2n'\bar{x}.\bar{x'}+n'\bar{x'}^2+\{x_n^2-2x_n\bar{x}+\bar{x}^2\}}{n'}} \quad (4)$$

โดย S' คือค่าเบี่ยงเบนมาตรฐานของข้อมูล n' ค่า

จาก (4) พิจารณาในส่วนของ $n'\bar{x}^2-2n'\bar{x}.\bar{x'}+n'\bar{x'}^2+\{x_n^2-2x_n\bar{x}+\bar{x}^2\}$ และแทน $\bar{x'}$ ด้วย (IV), n ด้วย (III)

$$n'\bar{x}^2-2n'\bar{x}.\bar{x'}+n'\bar{x'}^2+\{x_n^2-2x_n\bar{x}+\bar{x}^2\}$$
$$= n'\bar{x}^2-2n'\bar{x}(\frac{\bar{x}.n-x_n}{n'})+n'(\frac{\bar{x}.n-x_n}{n'})^2+\{x_n^2-2x_n\bar{x}+\bar{x}^2\}$$
$$= n'\bar{x}^2-2n\bar{x}^2+2x_n\bar{x}+n'(\frac{n^2\bar{x}^2-2n.\bar{x}.x_n+x_n^2}{n'^2})+\{x_n^2-2x_n\bar{x}+\bar{x}^2\}$$

$$= n'\bar{x}^2 - 2(n'+1)\bar{x}^2 + (\frac{(n'+1)^2\bar{x}^2 - 2(n'+1)\bar{x}.x_n + x_n^2}{n'}) + n'\{\frac{x_n^2 + \bar{x}^2}{n'}\}$$

$$= \frac{n'(n'\bar{x}^2 - 2n'\bar{x}^2 - 2\bar{x}^2)}{n'} + (\frac{n'^2\bar{x}^2 + 2n'\bar{x}^2 + \bar{x}^2 - 2n'x_n\bar{x} - 2\bar{x}.x_n + x_n^2}{n'}) + n'\{\frac{x_n^2 + \bar{x}^2}{n'}\}$$

$$= \frac{n'^2\bar{x}^2 - 2n'^2\bar{x}^2 - 2n'\bar{x}^2 + n'^2\bar{x}^2 + 2n'\bar{x}^2 + \bar{x}^2 - 2n'x_n\bar{x} - 2\bar{x}.x_n + x_n^2 + n'x_n^2 + n'\bar{x}^2}{n'}$$

$$= \frac{(n'+1)x_n^2 - 2(n'+1)x_n\bar{x} + (1+n')\bar{x}^2}{n'}$$

$$= \frac{(n'+1)(x_n^2 - 2x_n\bar{x} + \bar{x}^2)}{n'}$$

$$= \frac{n.(x_n - \bar{x})^2}{n'} \quad (5)$$

แทน (5) ลงใน (4) จะได้สมการหาค่าเบี่ยงเบนมาตรฐานในรูปแบบคำนวณต่อเนื่องของข้อมูลกลุ่มตัวอย่างดังนี้

$$S_n = \sqrt{\left(1 - \frac{1}{n\prime}\right) S'^2 + n\left(\frac{x_n - \bar{x}}{n\prime}\right)^2} \quad [2]$$

เมื่อ n คือ ขนาดข้อมูล (ครั้งหลังสุด) โดย $n>2$

n' คือ ขนาดข้อมูลครั้งก่อน โดย $n = n'+1$

x_n คือ ค่าสังเกตลำดับที่ n

S' คือ ค่าเบี่ยงเบนมาตรฐานครั้งก่อน

2. สมการหาค่าเบี่ยงเบนมาตรฐานแบบขยายต่อเนื่องสำหรับข้อมูลประชากร

เช่นเดียวกับการคำนวณหาค่าเบี่ยงเบนมาตรฐานของกลุ่มข้อมูลตัวอย่าง ถ้า x_i คือค่าสังเกตของกลุ่มข้อมูลประชากร ที่มีขนาดข้อมูลจำกัดทั้งหมด N ค่า การคำนวณหาค่าเบี่ยงเบนมาตรฐานของประชากร สามารถคำนวณได้จากสมการมาตรฐานทั่วไปดังนี้

ค่าเบี่ยงเบนมาตรฐาน $$\sigma_N = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}} \quad [3]$$

ซึ่งสามารถเขียนได้ดังนี้ $$\sigma_N = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_{n'}-\mu)^2 + (x_N-\mu)^2}{N}} \quad (6)$$

เมื่อพิจารณาเฉพาะส่วน $(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_{n'}-\mu)^2$ จะได้ดังนี้

$$(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_{n'}-\mu)^2$$

$$= (x_1^2 - 2x_1\mu + \mu^2) + (x_2^2 - 2x_2\mu + \mu^2) + ... + (x_{n'}^2 - 2x_{n'}\mu + \mu^2)$$

$$= (x_1^2 - 2x_1\mu' + \mu'^2) + (x_2^2 - 2x_2\mu' + \mu'^2) + ... + (x_{n'}^2 - 2x_{n'}\mu' + \mu'^2) +$$

$$\{(-2x_1\mu+\mu^2)+(-2x_2\mu+\mu^2)+\ldots+(-2x_{n'}\mu+\mu^2)\}-$$
$$\{(-2x_1\mu'+\mu'^2)+(-2x_2\mu'+\mu'^2)+\ldots+(-2x_{n'}\mu'+\mu'^2)\}$$
$$= (N-1)\sigma'^2+\{n'\mu^2-2\mu(x_1+x_2+..x_{n'})\} - \{n'\mu'^2-2\mu'(x_1+x_2+..x_{n'})\}$$
$$= (N-1)\sigma'^2 + \{n'\mu^2-2\mu.n'\mu'\} - \{n'\mu'^2-2\mu'n'\mu'\}$$
$$= (N-1)\sigma'^2 + n'\mu^2-2n'\mu.\mu'+n'\mu'^2 \qquad (7)$$

โดย $N > 2$ เหตุผลเช่นเดียวกับสมการ (V)

แทน (7) ลงใน (6)

$$\text{จะได้ } \sigma_N = \sqrt{\frac{(N-1)\sigma'^2+n'(\mu-\mu')^2+(x_N-\mu)^2}{N}} \qquad (8)$$

จาก (8) พิจารณาเฉพาะส่วน $n'(\mu-\mu')^2+(x_N-\mu)^2 = (n'\mu^2-2n'\mu.\mu'+n'\mu'^2)+(x_N^2-2x_N\mu+\mu^2)$

และแทน μ' ด้วยรูปแบบเช่นเดียวกับ (IV) และ N แทนด้วย (III)

$$n'(\mu-\mu')^2+(x_N-\mu)^2$$
$$= \{n'\mu^2-2n'\mu\frac{(\mu.N-x_N)}{n'}+n'(\frac{\mu.N-x_N}{n'})^2\}+(x_N{}^2-2x_N\mu+\mu^2)$$
$$= \{n'\mu^2-2N\mu^2+2\mu.x_N+(\frac{\mu^2N^2-2N\mu.x_N+x_N{}^2}{n'})\}+(x_N{}^2-2x_N\mu+\mu^2)$$
$$= \{n'\mu^2-2(n'+1)\mu^2+(\frac{\mu^2(n'+1)^2-2(n'+1)\mu.x_N+x_N{}^2}{n'})\}+(x_N{}^2+\mu^2)$$
$$= \frac{n'(n'\mu^2-2n'\mu^2-2\mu^2)}{n'}+(\frac{n'^2\mu^2+2n'\mu^2+\mu^2-2n'\mu.x_N-2\mu.x_N+x_N{}^2}{n'})+\frac{n'(x_N{}^2+\mu^2)}{n'}$$
$$=$$
$$\frac{n'^2\mu^2-2n'^2\mu^2-2n'\mu^2+n'^2\mu^2+2n'\mu^2+\mu^2-2n'\mu.x_N-2\mu.x_N+x_N{}^2+n'x_N{}^2+n'\mu^2}{n'}$$
$$= \frac{(n'+1)\mu^2-2(n'+1)x_N\mu+(n'+1)x_N{}^2}{n'} = \frac{N(x_N-\mu)^2}{n'} \qquad (9)$$

แทน (9) ลงใน (8)

$$\sigma_N = \sqrt{\frac{(N-1)\sigma'^2}{N}+\frac{\frac{N(x_N-\mu)^2}{n'}}{N}}$$ โดย $n' = N - 1$ หรือถ้ากำหนดให้ $N' = N - 1$

ในที่สุดจะได้สมการสำหรับการคำนวณค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่องของกลุ่มข้อมูลประชากรดังนี้

$$\sigma_N = \sqrt{\left(1-\frac{1}{N}\right)\sigma'^2+\frac{(x_N-\mu)^2}{N'}} \qquad [4]$$

โดย N คือ ขนาดข้อมูล (ครั้งหลังสุด) โดย $N > 2$

N' คือ ขนาดข้อมูลครั้งก่อน โดย $N = N'+1$

x_N คือ ค่าสังเกตลำดับที่ N

σ' คือ ค่าเบี่ยงเบนมาตรฐานครั้งก่อน

3. การตรวจสอบความถูกต้อง

จากสมการที่สร้างขึ้น ขั้นตอนต่อไปคือการนำสมการมาตรวจสอบความถูกต้อง โดยการหาค่าเบี่ยงเบนมาตรฐานจากข้อมูลที่สุ่มมา 50 ค่า เปรียบเทียบกับค่าเบี่ยงเบนมาตรฐานที่ได้จากสมการมาตรฐานทั่วไป สำหรับข้อมูลค่าแรกจะยังไม่มีค่าใด ๆ การหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานเริ่มต้น (S_0) จะเริ่มที่ข้อมูลค่าที่ 2 หรือ n=2 โดยอาศัยสมการพื้นฐาน (I): $\bar{x} = (x_1 + x_2)/2$ และ $S_0^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2$ ซึ่งการคำนวณหาค่าเบี่ยงเบนมาตรฐานค่าต่าง ๆ ของทั้งสองรูปแบบได้ผลดังนี้

ตารางที่ 1 ผลการตรวจสอบความถูกต้อง

#	X	X-Bar	*SD-ตัวอย่าง	**SD-ประชากร	จากสมการ [2]	จากสมการ [4]
1	780	780.00	-	-	-	-
2	450	615.00	233.345	165.000	* 233.345	** 165.000
3	90	440.00	345.109	281.780	345.109	281.780
4	490	452.50	282.887	244.987	282.887	244.987
5	815	525.00	293.769	262.755	293.769	262.755
6	499	520.67	262.969	240.057	262.969	240.057
7	160	469.14	276.062	255.584	276.062	255.584
8	240	440.50	268.116	250.800	268.116	250.800
9	570	454.89	254.487	239.933	254.487	239.933
10	730	482.40	255.219	242.122	255.219	242.122
11	550	488.55	242.978	231.670	242.978	231.670
12	834	517.33	252.222	241.484	252.222	241.484
13	119	486.69	265.556	255.138	265.556	255.138
14	800	509.07	268.527	258.760	268.527	258.760
15	440	504.47	259.373	250.578	259.373	250.578
16	790	522.31	260.548	252.274	260.548	252.274
17	808	539.12	261.617	253.806	261.617	253.806
18	588	541.83	254.067	246.909	254.067	246.909
19	170	522.26	261.229	254.262	261.229	254.262
20	166	504.45	266.449	259.703	266.449	259.703
21	180	489.00	269.181	262.694	269.181	262.694
22	798	503.05	270.828	264.601	270.828	264.601

* จากสมการ [1], ** จากสมการ [3]

ตารางที่ 1 ผลการตรวจสอบความถูกต้อง (ต่อ)

#	X	X-Bar	*SD-ตัวอย่าง	**SD-ประชากร	จากสมการ[2]	จากสมการ[4]
23	680	510.74	267.162	261.289	267.162	261.289
24	110	494.04	273.795	268.030	273.795	268.030
25	145	480.08	276.971	271.375	276.971	271.375
26	828	493.46	279.822	274.388	279.822	274.388
27	95	478.70	284.902	279.576	284.902	279.576
28	250	470.54	282.898	277.800	282.898	277.800
29	115	458.28	285.537	280.571	285.537	280.571
30	480	459.00	280.599	275.883	280.599	275.883
31	98	447.35	283.399	278.791	283.399	278.791
32	302	442.81	279.973	275.563	279.973	275.563
33	526	445.33	275.944	271.730	275.944	271.730
34	565	448.85	272.504	268.467	272.504	268.467
35	140	440.03	273.496	269.560	273.496	269.560
36	615	444.89	271.133	267.341	271.133	267.341
37	201	438.30	270.331	266.653	270.331	266.653
38	820	448.34	273.748	270.122	273.748	270.122
39	120	439.92	275.191	271.640	275.191	271.640
40	580	443.43	272.541	269.113	272.541	269.113
41	130	435.78	273.529	270.172	273.529	270.172
42	850	445.64	277.630	274.305	277.630	274.305
43	108	437.79	279.095	275.831	279.095	275.831
44	782	445.61	280.670	277.462	280.670	277.462
45	515	447.16	277.655	274.552	277.655	274.552
46	410	446.35	274.607	271.606	274.607	271.606
47	840	454.72	277.609	274.640	277.609	274.640
48	415	453.90	274.700	271.823	274.700	271.823
49	340	451.57	272.310	269.517	272.310	269.517
50	505	452.64	269.623	266.913	269.623	266.913

* จากสมการ [1], ** จากสมการ [3]

ผลการตรวจสอบ

จากตารางจะเห็นว่า การคำนวณหาค่าเบี่ยงเบนมาตรฐานโดยใช้สมการมาตรฐานทั่วไป เทียบกับค่าที่ได้จากสมการ ที่สร้างขึ้น จะให้คำตอบเหมือนกัน ทั้งในการคำนวณโดยใช้สมการสำหรับกลุ่มข้อมูลตัวอย่าง และสมการสำหรับกลุ่มข้อมูลประชากร และเป็นที่น่าสังเกตเมื่อขนาดข้อมูลมากขึ้น ค่าเบี่ยงเบนมาตรฐานที่คำนวณได้

จากสมการ ทั้งสองรูปแบบ ต่างมีลักษณะลู่เข้าหากันเช่นเดียวกับค่าที่ได้จากสมการมาตรฐานทั่วไป ซึ่งเป็นไปตามคุณสมบัติของค่าเบี่ยงเบนมาตรฐาน (กัลยา, 2553)

4. อัลกอริธึมสำหรับการคำนวณหาค่าเบี่ยงเบนมาตรฐาน

อัลกอริธึมการคำนวณหาค่าเบี่ยงเบนมาตรฐานสำหรับการใช้สมการมาตรฐานทั่วไป มีชื่อเรียกว่าอัลกอริธึมแบบ 2 ส่วน (two-pass algorithm) และอัลกอริธึมที่ใช้สมการแบบรวดเร็วต่อเนื่องเรียกว่า ออนไลน์-อัลกอริธึม (on-line algorithm) (สืบค้นจาก http://www.answers.com/topic/algorithms-for-calculating-variance และ http://en.wikipedia.org/wiki/Standard_deviation) ซึ่งมีรายละเอียดดังนี้

กำหนดให้ data[N] คือ ตัวแปรเก็บอาร์เรย์ข้อมูล ซึ่งสามารถมีค่ามากที่สุดไม่เกิน N ค่า

n คือ ตัวแปรเก็บขนาดข้อมูลปัจจุบัน โดย $n \in N$

n' คือ ตัวแปรเก็บขนาดข้อมูลครั้งก่อน โดย $n' = n - 1$

SD' คือ ตัวแปรเก็บค่าเบี่ยงเบนมาตรฐานครั้งก่อน

SD คือ ตัวแปรเก็บค่าเบี่ยงเบนมาตรฐานปัจจุบัน

x_n คือ ตัวแปรเก็บข้อมูลลำดับที่ n

AVR' คือ ตัวแปรเก็บค่าเฉลี่ยครั้งก่อน

AVR คือ ตัวแปรเก็บค่าเฉลี่ยปัจจุบัน

<u>อัลกอริธึมแบบ 2 ส่วน</u>

```
TowPart_STD (data[ ], n)
{   Sum = 0.0;
    Loop (i = 1 to n)
          Sum = Sum + data[i];
    AVR = Sum/n;
    Sum = 0.0;
    Loop (i = 1 to n)
          Sum = Sum + (data[i] - AVR)^2;
    SD = Sqrt(Sum/n);
    Return SD;
}
//โดย n = 2, 3, 4, ..., N
```

<u>อัลกอริธึมแบบขยายต่อเนื่อง</u>

```
Rapid_STD (Xn, n, SD′, AVR′)
{       n′ = n -1;
        AVR = (AVR′*n′ + Xn )/n;
        SD = Sqrt{(1-1/n)*SD′^2 + (Xn-AVR)^2/n′};
        Return SD, AVR;
}
//โดย n = 3, 4, ..., N
```

ส่วนอัลกอริธึมหลักสำหรับเรียกใช้ (call function) สามารถเขียนได้ดังนี้

```
ModuleCall( )
{   n = 1;
    Input(data[n]);
    While(not end of data)
          {  n = n+1;
             Input(data[n]);
             SD = TowPart_STD(data, n);
          }
    Return SD;
}
```

```
ModuleCall( )
{      n = 2;
       Input(X1, X2);
       AVR = (X1 + X2)/2;
       SD = Sqrt{((X1-AVR)^2+( X2-AVR)^2)/2};
       While(not end of data)
       { n = n+1;
         Input(Xn);
         (SD, AVR) = Rapid_STD (Xn, n, SD, AVR);
       }
       Return SD;
}
```

จากอัลกอริธึมทั้งสองวิธี จะเห็นว่าถ้าจำเป็นต้องหาค่าเบี่ยงเบนมาตรฐานทุกครั้ง เมื่อกลุ่มข้อมูลมีค่าสังเกตเพิ่มขึ้น โดยไม่สามารถคาดเดาจำนวนค่าสังเกตที่เพิ่มขึ้นได้ การหาค่าเบี่ยงเบนมาตรฐานด้วยอัลกอริธึมแบบ 2 ส่วน จะมีลักษณะในการทำงานแบบซ้อนลูป หมายความว่าถ้ามีค่าสังเกต 100 ข้อมูล (n=100) จะต้องทำซ้ำวนลูป 2x2 + 2x3 + 2x4 + ... + 2x100 = 100(100+1)-2 ครั้ง หรือ $n^2 + n - 2$ ครั้ง ซึ่งประสิทธิภาพของอัลกอริธึมวิธีนี้ จึงเป็นลักษณะของบิ๊กโอ-เอ็นกำลังสอง $O(n^2)$ ในขณะที่อัลกอริธึม แบบรวดเร็วต่อเนื่อง จะขึ้นอยู่กับลูปเพียงชั้นเดียว ดังนั้นการทำงานของ อัลกอริธึมวิธีนี้ จึงมีลักษณะเป็นแบบ บิ๊กโอ-เอ็น $O(n)$ ซึ่งถือว่ามีประสิทธิภาพมากกว่ากันมาก

อภิปรายและสรุปผลการดำเนินงาน

จากผลการดำเนินงาน สามารถสร้างสมการสำหรับการคำนวณค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่องที่ให้คำตอบถูกต้อง สำหรับรองรับการหาค่าเบี่ยงเบนมาตรฐาน ของกลุ่มข้อมูลตัวอย่าง และกลุ่มข้อมูลประชากร ในอีกรูปแบบหนึ่งที่ไม่ซับซ้อน เข้าใจง่าย และสะดวกในการใช้งาน โดยสมการทั้งสองจะขึ้นอยู่กับค่าเบี่ยงเบนมาตรฐานของค่าสังเกตก่อนหน้านี้ (S') ค่าสังเกตที่เพิ่มมาใหม่ (x_n) ขนาดข้อมูลปัจจุบัน (n) (รวมถึงค่าสังเกตที่เพิ่มมาใหม่) ค่าเฉลี่ยของค่าสังเกตปัจจุบัน ($\bar{x}_n$) หรือ $S_n(S', x_n, n, \bar{x}_n)$ เมื่อ $n > 2$ โดยค่าของ S' ไม่สามารถหาค่าได้เมื่อ n =1, 2 และที่ n=2 ค่าเบี่ยงเบนมาตรฐานเริ่มต้น (S_o) สามารถคำนวณได้จากสมการมาตรฐาน $S_0^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2$ หรือสำหรับกลุ่มข้อมูลประชากร $\sigma_o^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2}{2}$ ซึ่งในที่สุดสามารถสรุปรายละเอียดของสมการสำหรับการคำนวณค่าเบี่ยงเบนมาตรฐานแบบรวดเร็วต่อเนื่องได้ดังนี้

สำหรับกลุ่มตัวอย่าง

$$S_n^2 = \begin{cases} (1-\frac{1}{n'})s'^2 + n(\frac{x_n - \bar{x}}{n'})^2 & \text{if } n > 2;\ n' = n-1 \text{ and } S' \equiv S_{n-1} \\ S_o^2 & \text{otherwise} \end{cases}$$

สำหรับประชากร

$$\sigma_N^2 = \begin{cases} (1-\frac{1}{N})\sigma'^2 + \frac{(x_N - \mu)^2}{N'} & \text{if } N > 2;\ N' = N-1 \text{ and } \sigma' \equiv \sigma_{N-1} \\ \sigma_o^2 & \text{otherwise} \end{cases}$$

กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาการสารสนเทศที่ให้ความสนับสนุนมาโดยตลอด ขอขอบคุณ รศ.ดร.อำพล ธรรมเจริญ ดร.ณัฐนนท์ ลีลาตระกูล ที่ให้คำปรึกษาชี้แนะ และขอขอบคุณ รศ.ดร.สุกัญญา บูรณเดชาชัย ที่ให้คำชี้แนะและตรวจทาน

เอกสารอ้างอิง

กัลยา วานิชย์บัญชา. (2553). หลักสถิติ. พิมพ์ครั้งที่ 12. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย. 489 หน้า.

พยุง มีสัจ. (2553). Fuzzy System and Newral Networks: ILFN (Incremental Learning Fuzzy Neural Network). Chapter 7. http://suanpalm3.kmutnb.ac.th/teacher/phayung/.

Algorithms for Calculating Variance. http://www.answers.com/topic/algorithms-for-calculating-variance.

Bartz, A. E. (1999). Basic Statistical Concepts. 4th ed. Prentice-Hall, Inc. Simon & Schuster/A Viacon Company, Upper Saddle River, New Jersey 07458. 456 P.

Bruning, J. L. and Kintz, B. L. (1997). Computational Handbook of Statistics. 4th ed., One Jacob way, New York: Addison-Wesley Education Publicshers Inc.. 361 P.

Constructing a Fuzzy Expert System using the ILFN Network and The Genetic Algorithm. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=801205.

Navidi, W.. (2011). Statistics for Engineers and Scientists. 3rd ed. 1221 Avenue of The Americas, New York,: The McGraw-Hill Companies, Inc. 808 P.

Standard Deviation. http://en.wikipedia.org/wiki/Standard_deviation.

❑❑❑❑❑