តម្លៃអច្ឆន្តបរមាក្នុងសំណុំទិន្នន័យមួយអថេរ

Outliers in Univariate Data

ម៉ុង ម៉ាក

វិទ្យាស្ថានវិទ្យាសាស្ត្រនិងបច្ចេកវិទ្យា, រាជបណ្ឌិត្យសភាកម្ពុជា

ខែធ្នូ ឆ្នាំ២០២០

អត្ថបទនេះត្រូវបានត្រួតពិនិត្យដោយលោក **យឹម អាយុវឌ្ឍនៈវិជ្ជា** តំណាងនាយកដ្ឋានគណិតវិទ្យានិងស្ថិតិ

វិទ្យាស្ថានវិទ្យាសាស្ត្រនិងបច្ចេកវិទ្យា, រាជបណ្ឌិត្យសភាកម្ពុជា

មូលន័យសង្ខេប

ការដែលសំណុំទិន្នន័យផ្ទុកនូវតម្លៃអច្ចន្តបរមា វាជាបញ្ហាមួយនៅក្នុងការវិភាគទិន្នន័យស្ថិតិ។ អត្ថបទនេះនឹងពន្យល់អំពីអ្វីទៅដែលហៅថាតម្លៃអច្ចន្តបរមា, លើកឡើងអំពីវិធីមួយចំនួនសម្រាប់ពិនិត្យរកមើលតម្លៃទាំងនោះនៅក្នុងសំណុំទិន្នន័យ និង រកដំណោះស្រាយជាការឆ្លើយតបចំពោះវត្តមានរបស់វា។

តម្លៃអច្ឆន្តបរមាក្នុងសំណុំទិន្នន័យមួយអថេរ

សេចក្តីផ្តើម

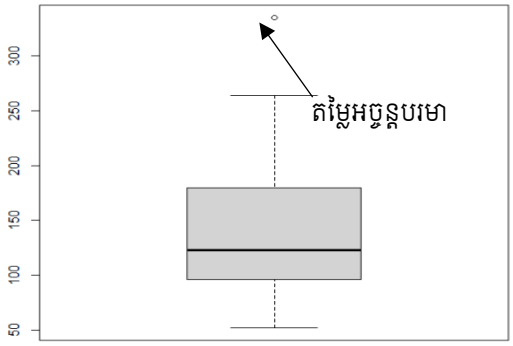
នៅក្នុងដំណើរការនៃការសិក្សាស្រាវជ្រាវ ការប្រមូលទិន្នន័យ ការបញ្ចូលទិន្នន័យដើម្បីវិភាគវាយតម្លៃនិងទាញការសន្និដ្ឋានគឺជាដំហានដ៏មានសារៈសំខាន់និងត្រូវធ្វើឡើងដោយម៉ត់ចត់។ វិធីវិភាគទិន្នន័យអថេរបរិមាណតម្រូវឱ្យមានការយកចិត្តទុកដាក់ខ្ពស់ចំពោះវត្តមាននៃតម្លៃអច្ឆន្តបរមានិងធ្វើការឆ្លើយតបដោយដំណោះស្រាយដាក់លាក់មួយ បើពុំនោះសោតទេ លទ្ធផលនៃការវិភាគនឹងមិនមានភាពត្រឹមត្រូវគ្រប់គ្រាន់និងជឿជាក់បានឡើយ។

តើអ្វីជាតម្លៃអច្ឆន្តបរមា? តើវាកើតឡើងដោយរបៀបណា? តើត្រូវដោះស្រាយបែបណាចំពោះវត្តមានរបស់វានៅក្នុងសំណុំទិន្នន័យ?

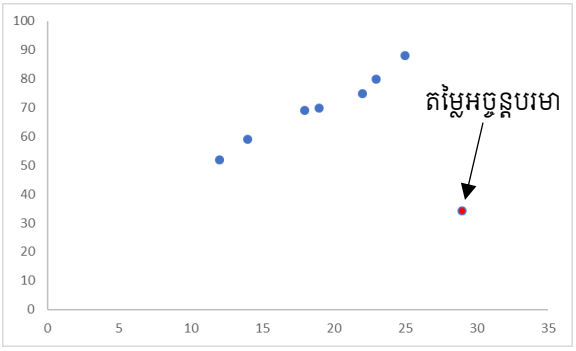
១. តម្លៃអច្ឆន្តបរមានិងការពិនិត្យរកមើលខា

នៅក្នុងស្ថិតិវិទ្យា «តម្លៃអច្ឆន្តបរមា»ជាពាក្យសម្រាប់ប្រើឱ្យត្រូវគ្នានឹងពាក្យ "outlier" ក្នុងភាសាអង់គ្លេសនៅក្នុងបរិបទស្ថិតិវិទ្យា។ ក្នុងន័យទូលាយ តម្លៃអច្ឆន្តបរមាសំដៅទៅលើតម្លៃដែលតូចខ្លាំងឬធំខ្លាំងហួសហេតុបើធៀបទៅនឹងតម្លៃដទៃទៀតនៅក្នុងសំណុំទិន្នន័យ[1] (Bluman, 2014, p. 121; Grubbs, 1969; Maddala, 1992, p. 89) ។

តម្លៃអច្ឆន្តបរមាអាចមាននៅក្នុងសំណុំទិន្នន័យមួយអថេរ(univariate data)(រូបទី១)ដែលគេហៅវាថា តម្លៃអច្ឆន្តបរមាមួយអថេរ(univariate outlier) ឬនៅក្នុងសំណុំទិន្នន័យពហុអថេរ(multivariate data)ដែលគេហៅថា តម្លៃអច្ឆន្តបរមាពហុអថេរ (multivariate outlier) (រូបទី២) (DataVedas, 2018, February 2; RAY, 2016, January 10; "Univariate and Multivariate Outliers," 2020, June 24)។

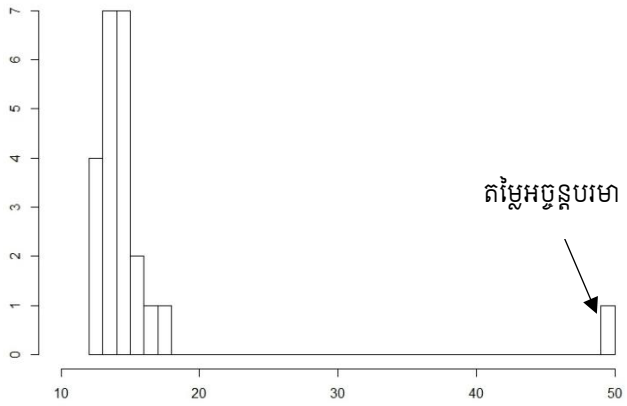


រូបទី១ Boxplot បង្ហាញតម្លៃអច្ឆន្តបរមាមួយអថេរ



រូបទី២ Scatter plot បង្ហាញតម្លៃអច្ឆន្តបរមាពីរអថេរ

[1] ដែលជាប្រភេទសំណុំទិន្នន័យបរិមាណ(quantitative data)



រូបទី៣ អ៊ីស្តូក្រាម បង្ហាញតម្លៃអច្ឆន្តបរមា

ក្រៅអំពីការពិនិត្យរកមើលតាមរយៈក្រាហ្វ(រូបទី១, រូបទី២ និង រូបទី៣) តម្លៃអច្ឆន្តបរមាអាចត្រូវពិនិត្យរកមើលដោយប្រើប្រាស់វិធីមួយចំនួនដែលក្នុងនោះរួមមានតម្លៃជាវង្វាស់ជាក់លាក់ឬតេស្តស្ថិតិ។

១.១. **វិធី** Tukey's Fence

ការពិនិត្យមើលជាបឋមទៅលើវង្វាស់ស្ថិតិពិពណ៌នាដូចជាតម្លៃធំបំផុត និង តូចបំផុតជាដើមនោះអាចឱ្យគេជាក់ការសង្ស័យអំពីវត្តមាននៃតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យបាន។ ដើម្បីឱ្យបានជាក់លាក់ជាងនេះ គេអាចប្រើវិធីមួយហៅថា Tukey's Fence ដោយផ្អែកលើកាទីល និង រ៉ង់អាំងទែកាទីល។ វិធីនេះត្រូវបានប្រើប្រាស់ជាទូទៅ។ តាមរយៈវិធីនេះ តម្លៃទាំងឡាយណាដែលតូចជាង $Q_1 - 1.5 \cdot IQR$ ឬធំជាង $Q_3 + 1.5 \cdot IQR$ ត្រូវចាត់ទុកជាតម្លៃអច្ឆន្តបរមាដែល $IQR = Q_3 - Q_1$ ហៅថារ៉ង់អាំងទែកាទីល។ ឧទាហរណ៍ ចំពោះសំណុំទិន្នន័យ

15, 14, 15, 12, 16, 14, 15, 13, 15, 14, 13, 14, 15, 16, 13, 15, 14, 15, 14, 17, 14, 18, 50

កាទីលទី១និងទី៣គឺ $Q_1 = 14$ និង $Q_3 = 15$។ ដូច្នេះរ៉ង់អាំងទែកាទីលគឺ $IQR = 15 - 14 = 1$ ។

$$Q_1 - 1.5 \cdot IQR = 14 - 1.5 \cdot 1 = 12.5$$

$$Q_3 + 1.5 \cdot IQR = 15 + 1.5 \cdot 1 = 16.5$$

ដូច្នេះតម្លៃ 12, 17, 18 និង 50 ត្រូវចាត់ទុកជាតម្លៃអច្ឆន្តបរមា។

១.២. **ការប្រើតម្លៃ** z (z-scores) **និង តម្លៃ** z **កែលម្អ** (Modified z-score)

តម្លៃ z(z-score) នៃតម្លៃ xមួយនៅក្នុងសំណុំទិន្នន័យ គឺជាចំនួនគម្លាតស្តង់ដាដែលតម្លៃនោះស្ថិតនៅខាងលើមធ្យមឬនៅខាងក្រោមមធ្យមរបស់សំណុំទិន្នន័យ។ ចំពោះសំណុំទិន្នន័យដែលមានបំណែងចែកណរម៉ាល់ស្ទើរតែទាំងអស់ (99.7%) នៃតម្លៃក្នុងសំណុំទិន្នន័យ ស្ថិតនៅក្នុងចន្លោះ3គម្លាតស្តង់ដាពីមធ្យមរបស់សំណុំទិន្នន័យ

នោះ។ ដូច្នេះ តម្លៃទាំងឡាយណានៅក្នុងសំណុំទិន្នន័យណារម៉ាល់ ដែលមានតម្លៃ zរបស់វាតូចជាង - 3 ឬធំជាង + 3 គឺអាចចាត់ទុកជាតម្លៃអច្ឆន្តបរមា។ វិធីស្រដៀងគ្នានេះហៅថា modified z-score បង្ហាញដោយ Iglewicz and Hoaglin (1993) ដូចដែលត្រូវបានយោងនៅក្នុង NIST/SEMATECH (2003a) ក៏ដូចជា Kannan, Manoj, and Arumugam (2015) កំណត់ដោយ

$$M_i = \frac{0.6745\left(x_i - \tilde{x}\right)}{median\left(\left|x_i - \tilde{x}\right|\right)} \qquad (១)$$

ដែល $\tilde{x}$ គឺជាមេដ្យានរបស់សំណុំទិន្នន័យ។ ចំពោះ x_i ទាំងឡាយណាដែលធ្វើឱ្យ $|M_i| > 3.5$ ត្រូវសម្គាល់ថាជាតម្លៃអច្ឆន្តបរមា។

១.៣. គម្លាតដាច់ខាតនៃមេដ្យាន (Median Absolute Deviation)

វិធីតម្លៃz ទទួលរងឥទ្ធិពលខ្លាំងក្លាអំពីតម្លៃដែលតូច/ធំហួសហេតុ ដោយសារវាអាស្រ័យលើមធ្យម និងគម្លាតស្តង់ដារបស់សំណុំទិន្នន័យ។ ដូច្នេះជំនួសឱ្យវិធីតម្លៃ z គេប្រើប្រាស់វិធីមួយទៀតដែលផ្អែកលើមេដ្យានដែរនោះ គឺ វិធីគម្លាតដាច់ខាតនៃមេដ្យាន ដូចដែលមានលើកឡើងដោយ Kannan et al. (2015) Leys, Ley, Klein, Bernard, and Licata (2013) ជាដើម។ ដំណើរការនៃវិធីនេះគឺដំបូងគណនាគម្លាតដាច់ខាតនៃមេដ្យាន (MAD)

$$MAD = median\left(\left|x_i - \tilde{x}\right|\right) \qquad (២)$$

បន្ទាប់មក រកតម្លៃ

$$\tilde{x} \pm k \times 1.4826\, MAD \qquad (៣)$$

ដែល $k = 2, 2.5, 3$ ក្នុងនោះតម្លៃ $k = 3$ ជាភាពអភិរក្សនិយមខ្ពស់ $k = 2.5$ ជាភាពអភិរក្សនិយមមធ្យម និង $k = 2$ ជាភាពអភិរក្សនិយមទាប នេះបើតាម (Miller, 1991) ដូចដែលមានយោងនៅក្នុង (Kannan et al., 2015)។ តម្លៃទាំងឡាយណាដែលស្ថិតនៅក្រៅចន្លោះដែលកំណត់ដោយកន្សោមខាងលើ ត្រូវសម្គាល់ថាជាតម្លៃអច្ឆន្តបរមា។

ក្រៅអំពីវិធីដែលបានបរិយាយខាងលើ នៅមានវិធីមួយចំនួនទៀតដែលផ្អែកលើតេស្តសម្មតិកម្មស្ថិតិសម្រាប់រកមើលតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យ។ ខាងក្រោមនេះ គឺជាតេស្តសម្មតិកម្មស្ថិតិមួយចំនួនដែលគេនិយមប្រើប្រាស់៖

១.៤. តេស្ត Grubbs (Grubbs' Test)

តេស្ត Grubbs (Grubbs, 1969; Stefansky, 1972) អាចត្រូវប្រើប្រាស់ដើម្បីរកមើលតម្លៃអច្ឆន្តបរមានៅ

ក្នុងសំណុំទិន្នន័យមួយអថេរដែលមានរបាយណរម៉ាល់ឬប្រហាក់ប្រហែលណរម៉ាល់ (NIST/SEMATECH, 2003a)។ សម្មតិកម្មសម្រាប់តេស្តកំណត់ដោយ៖

H_0: គ្មានតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យ

H_1: មានតម្លៃអច្ឆន្តបរមាចំនួនមួយនៅក្នុងសំណុំទិន្នន័យ

តម្លៃស្ថិតិតេស្តសម្រាប់តេស្តសងខាងកំណត់ដោយ

$$G = \frac{\max|x_i - \bar{x}|}{s} \qquad (៤)$$

ដែល $\bar{x}$ គឺជាមធ្យមគំរូតាង និង s គឺជាគម្លាតស្តង់ដាគំរូតាង។ ចំពោះតេស្តសងខាង សម្មតិកម្ម H_0 ត្រូវបដិសេធ និងសន្និដ្ឋានថាមានតម្លៃអច្ឆន្តបរមាចំនួនមួយនៅក្នុងសំណុំទិន្នន័យ បើ

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{\left(t_{\alpha/2N,N-2}\right)^2}{N-2+\left(t_{\alpha/2N,N-2}\right)^2}} \qquad (៥)$$

ដែល $t_{\alpha/2N,N-2}$ តាងឱ្យតម្លៃវិនិច្ឆ័យនិងរកបានពីបំណែងចែកស្ទូដិន (Student's distribution) ចំពោះដឺក្រេសេរី $N-2$ និង កម្រិតសារៈសំខាន់ $\alpha/2N$ [2] [3]។

ចំពោះតេស្តម្ខាង តម្លៃស្ថិតិតេស្តកំណត់ដោយ

- តេស្តមើលថាតើតម្លៃតូចបំផុតក្នុងសំណុំទិន្នន័យ ជាតម្លៃអច្ឆន្តបរមាឬទេ

$$G = \frac{\bar{x} - x_{\min}}{s} \qquad (៦)$$

ដែល $x_{\min}$ ជាតម្លៃតូចបំផុតនៅក្នុងគំរូតាង។

- តេស្តមើលថាតើតម្លៃធំបំផុតក្នុងសំណុំទិន្នន័យ ជាតម្លៃអច្ឆន្តបរមាឬទេ

$$G = \frac{x_{\max} - \bar{x}}{s} \qquad (៧)$$

ដែល $x_{\max}$ ជាតម្លៃធំបំផុតនៅក្នុងគំរូតាង។

នៅក្នុងតេស្តម្ខាង សម្មតិកម្ម H_0 ត្រូវបដិសេធ និង សន្និដ្ឋានថាមានតម្លៃអច្ឆន្តបរមាចំនួនមួយនៅក្នុងសំណុំទិន្នន័យ បើ

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{\left(t_{\alpha/N,N-2}\right)^2}{N-2+\left(t_{\alpha/N,N-2}\right)^2}} \qquad (៨)$$

---

[2] N គឺជាទំហំគំរូតាង

[3] គួរកត់សម្គាល់អំពីភាពខុសគ្នារវាងកម្រិតសារៈសំខាន់ α សម្រាប់តេស្តទាំងមូល និង កម្រិតសារៈសំខាន់សម្រាប់រកតម្លៃ t នៅក្នុងបំណែងចែកស្ទូដិន t។

តេស្តនេះអាចអនុវត្តនៅក្នុងសុសវែរ R ដោយអនុគន៍ grubbs.test (data, type = 10, opposite = FALSE, two.sided = FALSE) នៅក្នុង package ដែលមានឈ្មោះថា outliers ។ ប៉ារ៉ាម៉ែត្រ type = 10 ដែលជាតម្លៃកំណត់ស្រាប់ ប្រើសម្រាប់តេស្តពិនិត្យមើលតម្លៃអច្ឆន្តបរមាតែមួយគត់សំដៅទៅលើតម្លៃណាដែលមានគម្លាតធំជាងគេពីមធ្យមរបស់សំណុំទិន្នន័យ។ ដំណើរការអាចផ្តោតទៅរកកន្ទុយខាងស្តាំឬខាងឆ្វេងដោយស្វ័យប្រវត្តិ និង អាចកំណត់ឱ្យបញ្ច្រាសវិញបានតាមរយៈការផ្លាស់ប្តូរតម្លៃរបស់ប៉ារ៉ាម៉ែត្រ opposite។ ប៉ារ៉ាម៉ែត្រ type = 11 សម្រាប់តេស្តតម្លៃអច្ឆន្តបរមាចំនួនពីរនៅកន្ទុយទាំងសងខាង រីឯ type = 20 សម្រាប់តេស្តតម្លៃអច្ឆន្តបរមាចំនួនពីរនៅកន្ទុយតែម្ខាង។

១.៥. តេស្ត Dixon (Dixon's Test)

តេស្តនេះមានឈ្មោះផ្សេងទៀតដូចជា Dixon's Q-test ឬ Q test។ តេស្តនេះប្រើសម្រាប់ពិនិត្យរកមើលតម្លៃអច្ឆន្តបរមាចំនួនមួយ(តែមួយគត់)នៅក្នុងសំណុំទិន្នន័យដែលមានរបាយណារម៉ាល់ និង ទំហំគំរូតាងតូចដែលអាចពី 3 ទៅ 10 (NIST/SEMATECH, 2003b)។ បើតាម Soetewey (2020) ទំហំគំរូតាងសម្រាប់តេស្តនេះគឺ 25 ចុះក្រោម។

សម្មតិកម្មសម្រាប់តេស្តនេះអាចថ្លែងដូចខាងក្រោម

H_0 : គ្មានតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យ

H_1 : តម្លៃតូចបំផុត(ឬធំបំផុត) នៅក្នុងសំណុំទិន្នន័យជាតម្លៃអច្ឆន្តបរមា

តេស្តនេះអាចអនុវត្តនៅក្នុងសុសវែរ R តាមរយៈអនុគមន៍

dixon.test(data, type = 0, opposite = FALSE, two.sided = TRUE)

ចំពោះប៉ារ៉ាម៉ែត្ររបស់អនុគមន៍ ក្រៅពី type = 0 ដែលជាតម្លៃកំណត់ស្រាប់ វាមានជម្រើសផ្សេងៗទៀតដូចជា type = 10 សម្រាប់គំរូតាងទំហំ 3-7, type = 11 សម្រាប់គំរូតាងទំហំ 8-10, type = 21 សម្រាប់គំរូតាងទំហំ 11-13 និង type = 22 សម្រាប់គំរូតាងទំហំ 14 ឡើងទៅ។

១.៦. តេស្ត Tietjen-Moore (Tietjen-Moore test)

តេស្ត Tietjen-Moore (Tietjen & Moore, 1972) ដូចដែលយោងនៅក្នុង NIST/SEMATECH (2003d) អាចទុកជាការពង្រីកឱ្យកាន់តែមានភាពទូទៅបន្តពីតេស្ត Grubbs សម្រាប់ពិនិត្យរកមើលតម្លៃអច្ឆន្តបរមាចំនួនច្រើននៅក្នុងសំណុំទិន្នន័យមួយអថេរ។ តេស្តនេះតម្រូវឱ្យមានការកំណត់ដោយជាក់លាក់នូវចំនួន k នៃតម្លៃដែលសង្ស័យថាជាតម្លៃអច្ឆន្តបរមា។ សម្មតិកម្មសម្រាប់តេស្តនេះគឺ៖

H_0 : គ្មានតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យ

H_1 : មានតម្លៃអច្ឆន្តបរមាចំនួន k នៅក្នុងសំណុំទិន្នន័យ

តេស្ត Tietjen-Moore ជាប្រភេទតេស្តខាងឆ្វេងជានិច្ច។ ការគណនាតម្លៃស្ថិតិតេស្តនិងការរកតម្លៃនិច្ឆ័យ ទាមទារនូវកិច្ចការដែលប្រទាក់ក្រឡាច្រើន ដូច្នេះការធ្វើតេស្តអាចអនុវត្តនៅក្នុងសុសវែរ R។ អនុគមន៍សម្រាប់ តេស្តគឺ FindOutliersTietjenMooreTest (data, k, alpha = 0.05) នៅក្នុង package ដែលមានឈ្មោះថា climtrends (Gama, n.d.)។ ប៉ារ៉ាម៉ែត្រ k របស់អនុគមន៍ ជាចំនួនតម្លៃដែលគេសង្ស័យថាជាតម្លៃអច្ឆន្តបរមា ហើយ alpha ដែលមានតម្លៃកំណត់ស្រាប់ 0.05 គឺជាកម្រិតសារៈសំខាន់។

ធាតុចេញ (output) របស់អនុគមន៍មានពីរសំខាន់ ទីមួយគឺតម្លៃស្ថិតិតេស្ត និង មួយទៀតជាតម្លៃនិច្ឆ័យ របស់តេស្ត។ បើតម្លៃស្ថិតិតេស្តតូចជាងតម្លៃនិច្ឆ័យ នោះ H_0 នឹងត្រូវបដិសេធហើយគេសន្និដ្ឋានថា តម្លៃសង្ស័យទាំង នោះ ជាតម្លៃអច្ឆន្តបរមា។

១.៧. តេស្ត Generalized ESD (Generalized Extreme Studentized Deviate Test)

តេស្តនេះអាចប្រើប្រាស់ ដើម្បីពិនិត្យរកមើលតម្លៃអច្ឆន្តបរមាចំនួនមួយឬលើសពីមួយនៅក្នុងសំណុំទិន្នន័យ មួយអថេរដែលមានបំណែងចែកប្រហាក់ប្រហែលណរម៉ាល់។ តេស្ត Generalized ESD ទាមទារនូវការកំណត់តម្លៃ k ដែលតាងឱ្យចំនួនច្រើនបំផុតនៃតម្លៃដែលត្រូវសង្ស័យថាជាតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យ (Rosner, 1983) ដូចដែលមានយោងនៅក្នុង NIST/SEMATECH (2003c)។ ទំហំគំរូតាងដែលតេស្តធ្វើបានត្រឹមត្រូវបំផុត គឺ $n \geq 25$ ប៉ុន្តែ $n \geq 15$ ក៏អាចទទួលយកបាន។

H_0 : គ្មានតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យ

H_1 : មានតម្លៃអច្ឆន្តបរមារហូតដល់ចំនួនkនៅក្នុងសំណុំទិន្នន័យ

តេស្ត Generalized ESD អាចអនុវត្តបានក្នុងសុសវែរ R នៅក្នុង package ដែលមានឈ្មោះថា EnvStats តាមរយៈអនុគមន៍ rosnerTest(data, k=3,alpha=0.05,warn=TRUE)។ តម្លៃ k= 3 ជាតម្លៃកំណត់ស្រាប់ក្នុង អនុគមន៍ ប៉ុន្តែតម្លៃ k អាចកំណត់យកពី 1 ដល់ n – 2 ដែល n ជាចំនួនតម្លៃនៅក្នុងសំណុំទិន្នន័យ។ តម្លៃ alpha = 0.05 ជាកម្រិតសារៈសំខាន់ (ប្រូបាប៊ីលីតេនៃកំហុសប្រភេទ I) ដែលកំណត់ស្រាប់នៅក្នុងអនុគមន៍ ប៉ុន្តែអាច កំណត់យកតម្លៃផ្សេងទៀតនៅក្នុងចន្លោះ 0 និង 1។ ការកំណត់ស្រាប់ warn = TRUE គឺដើម្បីបង្ហាញសារព្រមាន នៅពេលដែលការកំណត់តម្លៃ alpha និង តម្លៃ k មិនសមស្របនឹងគ្នា។

២. ប្រភពនៃតម្លៃអច្ឆន្តបរមា

តម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យអាចកើតចេញមកពីហេតុមួយចំនួនដូចជា៖ កំហុសដោយមនុស្សឬ ដោយម៉ាស៊ីននៅពេលដែលបញ្ចូលនិង/ឬកត់ត្រាទិន្នន័យ(Morgan, 2016, April 15)បាតុភូតឬព្រឹត្តិការណ៍ផ្សេងៗ ឬ ធាតុនៃស្ថិតិសាកលផ្សេងត្រូវបញ្ចូលមកក្នុងគំរូតាងដែលគេសិក្សា (Blatná, 2006)។

ឧទាហរណ៍ខ្លះៗអំពីប្រភពដែលនាំឱ្យកើតមានតម្លៃអច្ចន្តបរមានៅក្នុងសំណុំទិន្នន័យ៖

១. ប្រាក់ខែរបស់បុគ្គលិកម្នាក់គឺ ១០០០ ដុល្លារ ប៉ុន្តែការបែញ្ចូលលើសលេខ ០ ចំនួនមួយឬពីរ នោះតម្លៃនៅក្នុងសំណុំទិន្នន័យនឹងទៅជា ១០០០០ ដុល្លារ ឬ ១០០០០០ ដុល្លារ ឬថា ការបែញ្ចូលទិន្នន័យបញ្ចូលខ្វះលេខ ០ ក៏នាំឱ្យតម្លៃបញ្ចូលខុសនោះអាចជាតម្លៃអច្ចន្តបរមាផងដែរ។

២. គេមានជញ្ជីងចំនួន ១០ ដែលក្នុងនោះជញ្ជីងចំនួន៩ មានភាពប្រក្រតី រីឯជញ្ជីងមួយទៀតមិនប្រក្រតី។ គេប្រើប្រាស់ជញ្ជីងទាំងនេះដើម្បីថ្លឹងផ្លែទុរេន។ ជញ្ជីងដែលមិនមានភាពប្រក្រតីអាចនឹងបង្ហាញរង្វាស់ទម្ងន់[4] ដែលមានតម្លៃតូចខ្លាំងឬធំខុសពីប្រក្រតី។ តម្លៃអច្ចន្តបរមាប្រភេទនេះកើតមកអំពីអ្វីដែលគេហៅថា ភាពល្អៀងរង្វាស់ (measurement error) (RAY, 2016, January 10)។

៣. នៅក្នុងការសម្គាសប្រមូលទិន្នន័យពីយុវជនមួយក្រុមអំពីចំនួនបារីដែលពួកគេជក់នៅក្នុង១ថ្ងៃ ដោយហេតុផលផ្សេងៗ យុវជនខ្លះមិនប្រាប់តួលេខពិតទេ ពួកគេអាចឆ្លើយដោយបន្ថយចំនួនមកនៅតិចតួច។ ចំនួនដែលពួកគេឆ្លើយកូតករទាំងនេះអាចជាចំនួនដែលតូចមិនប្រក្រតីបើធៀបទៅនឹងចំនួនដែលបានពីអ្នកឆ្លើយផ្សេងទៀត។ តម្លៃអច្ចន្តបរមាទាំងនេះហៅថា តម្លៃអច្ចន្តបរមាចេតនា (intentional outlier) (DataVedas, 2018, February 2; RAY, 2016, January 10)។

៤. កំហុសនៅក្នុងការប្រើប្រាស់ឯកតារង្វាស់ ដូចជា រវាង (គីឡូក្រាម និង ក្រាម គីឡូម៉ែត្រនិងសង់ទីម៉ែត្រជាដើម) ក៏ជាហេតុមួយនាំឱ្យមានតម្លៃអច្ចន្តបរមាផងដែរ (DataVedas, 2018, February 2)។

៥. នៅក្នុងការប្រមូលទិន្នន័យប្រាក់បៀវត្សរបស់គ្រូបង្រៀនស្រាប់តែដោយកំហុសណាមួយនោះ ប្រាក់បៀវត្សរបស់បុគ្គលិកការងារផ្សេងត្រូវរាប់បញ្ចូលដែរ។ តម្លៃអច្ចន្តបរមាបែបនេះកើតឡើងពីភាពល្អៀងមួយប្រភេទដែលគេហៅថា ភាពល្អៀងគំរូតាង (sampling error) (DataVedas, 2018, February 2; RAY, 2016, January 10)។

៦. តម្លៃអច្ចន្តបរមាមួយប្រភេទទៀតកើតឡើងតាមសភាពរបស់វា ពុំមែនដោយសារមនុស្សធ្វើឱ្យវាកើតឡើងនោះទេ ជាឧទាហរណ៍ ប្រាក់ចំណូលរបស់បុគ្គល។ បុគ្គលខ្លះមានប្រាក់ចំណូលយ៉ាងខ្ពស់ បុគ្គលខ្លះមានប្រាក់ចំណូលទាបក្រៃលែង។ តម្លៃប្រភេទនេះ ឈ្មោះថា តម្លៃអច្ចន្តបរមាធម្មតា (natural outlier) (DataVedas, 2018, February 2)។

៧. ការមិនអាចកំណត់ដឹងបាននូវតម្លៃដែលគេប្រើ ដើម្បីតាងឱ្យតម្លៃបាត់ (missing value) ("Univariate and Multivariate Outliers," 2020, June 24) ។

[4] នៅក្នុងភាសាទូទៅយើងនិយាយថា ថ្លឹងទម្ងន់។ នៅក្នុងរូបវិទ្យាគឺ ថ្លឹងម៉ាស់។

៣. ដំណោះស្រាយចំពោះតម្លៃអច្ចន្តបរមា

ផ្នែកទី១លើកឡើងអំពីវិធីមួយចំនួនសម្រាប់ពិនិត្យរកមើលតម្លៃអច្ចន្តបរមា។ ផ្នែកទី២លើកឡើងអំពីមូលហេតុដែលនាំឱ្យកើតមានតម្លៃអច្ចន្តបរមានៅក្នុងសំណុំទិន្នន័យ។ នៅក្នុងផ្នែកទី៣នេះ ដំណោះស្រាយមួយចំនួនចំពោះតម្លៃអច្ចន្តបរមានៅក្នុងសំណុំទិន្នន័យមួយអថេរនឹងត្រូវបង្ហាញតាមអ្វីដែលធ្លាប់ត្រូវបានគេអនុវត្ត។

១. តាមរបៀបសាមញ្ញ នៅពេលដែលពិនិត្យឃើញតម្លៃអច្ចន្តបរមានៅក្នុងសំណុំទិន្នន័យ គេអាចធ្វើការផ្ទៀងផ្ទាត់ការបញ្ចូលទិន្នន័យឡើងវិញ ពីព្រោះវាអាចទាក់ទងនឹងកំហុសជុំវិញការបញ្ចូលទិន្នន័យ ឬ ផ្ទៀងផ្ទាត់ឡើងវិញនូវការកត់ត្រាចម្លើយ ឬ ឧបករណ៍រង្វាស់នៅក្នុងការប្រមូលទិន្នន័យ។ នៅក្នុងករណីដែលគេមិនអាចកែតម្រូវបានហើយដឹងថាតម្លៃទាំងនេះ វាមិនអាចទទួលយកបានទាល់តែសោះ ដំណោះស្រាយគឺលុបវាចោល ជាពិសេសនៅក្នុងករណីដែលមានចំនួនតិចតួច។

២. នៅក្នុងករណីខ្លះ គេធ្វើការវិភាគនិងសរសេររបាយការណ៍បង្ហាញទាំងពីរករណីៈ ករណីដែលមានតម្លៃអច្ចន្តបរមានៅក្នុងនោះ និង ករណីដែលតម្លៃអច្ចន្តបរមាត្រូវបានដកចេញ។ ការធ្វើដូច្នេះជាការបង្ហាញអំពីតម្លាភាពមួយ ធ្វើឱ្យកិច្ចការទទួលបាននូវការជឿទុកចិត្ត។ នៅក្នុងករណីខ្លះទៀតដែលតម្លៃអច្ចន្តបរមាមិនធ្វើឱ្យលទ្ធផលរងការប៉ះពាល់ខ្លាំង ប៉ុន្តែវាគ្រាន់តែបំពានលក្ខខណ្ឌរបស់វិធីវិភាគ គេរក្សាទុកវាដដែលដោយគ្រាន់តែសរសេរបង្ហាញនៅក្នុងផ្នែកដែនកំណត់ (limitation) របស់អត្ថបទ។

៣. កាត់តម្រឹមទិន្នន័យយកតែរ៉ង់ដែលសមស្រប ឧទាហរណ៍ អាយុយកចន្លោះពី 0 ទៅ 100 ឬ ចន្លោះសតភាគទី5 ទៅសតភាគទី95 ជាដើម។

៤. ប្រើវិធីវីនសរ៍(Winsorizing or Winsorization method) ដើម្បីកាត់បន្ថយឥទ្ធិពលនៃតម្លៃអច្ចន្តបរមា។

៥. នៅក្នុងករណីដែលតម្លៃអច្ចន្តបរមាបង្កឱ្យមានអណរម៉ាល់ភាព (non-normality) គេអាចប្រើវិធីបំប្លែងអថេរមានដូចជា ការប្រើប្រាស់លោការីតជាដើម ដើម្បីបានទិន្នន័យណរម៉ាល់។

៦. ប្រើប្រាស់ robust statistics នៅក្នុងការវិភាគទិន្នន័យ។

៧. នៅក្នុងអថេរស៊េរីពេល តម្លៃដែលបាត់ (missing values) ឬ តម្លៃអច្ចន្តបរមាត្រូវបានដាក់ជំនួសមកវិញដោយតម្លៃដែលសមស្រប តាមរយៈវិធីមួយដែលហៅថា វិធីដាក់បញ្ចូល (imputation method) ។

តម្លៃអច្ឆន្តបរមាសំដៅដល់តម្លៃដែលធំហួសហេតុឬតូចហួសហេតុបើធៀបនឹងតម្លៃដទៃទៀតនៅក្នុងសំណុំទិន្នន័យ។ វាអាចកើតឡើងដោយកំហុស ឬ មិនមែនដោយសារកំហុសទេពោលគឺដោយសភាពដូច្នោះរបស់វានៅក្នុងសំណុំទិន្នន័យ។ សំណង់ក្រាហ្វ/ដ្យាក្រាមស្ថិតិ (ដូចជា boxplot, scatter plot និង អ៊ីស្តូក្រាម) និង វិធីស្ថិតិពិពណ៌នាជួយឱ្យគេអាចរកឃើញឬដាក់ការសង្ស័យអំពីវត្តមានតម្លៃអច្ឆន្តបរមា។ លើសពីនេះក៏នៅមានតេស្តសម្មតិកម្មស្ថិតិមួយចំនួនទៀតដែលអាចប្រើប្រាស់សម្រាប់រកមើលតម្លៃអច្ឆន្តបរមា។ បន្ទាប់ពីរកឃើញតម្លៃអច្ឆន្តបរមា ការឆ្លើយតបចំពោះវាគឺអាស្រ័យទៅតាមស្ថានភាព ដូចជាកែតម្រូវឡើងវិញដោយការផ្ទៀងផ្ទាត់ចម្លើយនៅក្នុងកម្រងសំណួរ ឬ កែតម្រូវដោយវិធី imputation។ បើមិនអាចកែបានទេ ត្រូវលុបចោល ឬ ត្រូវរក្សាទុកនិងប្រើវិធីវិភាគនិង/ឬសរសេរបាយការណ៍ដែលសមស្រប។

អត្ថបទនេះសិក្សាតែអំពីតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យមួយអថេរតែប៉ុណ្ណោះ។ ដូច្នេះអត្ថបទក្រោយទៀតគួរសិក្សាអំពីតម្លៃអច្ឆន្តបរមានៅក្នុងសំណុំទិន្នន័យច្រើនអថេរ៕

ឯកសារពិគ្រោះ

Blatná, D. (2006). Outliers in regression. *Trutnov, 30*, 1-6.

Bluman, A. (2014). *ELEMENTARY STATISTICS: A STEP BY STEP APPROACH*. USA: McGraw-Hill Higher Education.

DataVedas. (2018, February 2). OUTLIER TREATMENT. Retrieved from https://www.datavedas.com/outlier-treatment/

Gama, J. (n.d.). findOutliers.Tietjen.Moore.test: Find outliers based on the Tietjen Moore test. Retrieved from https://rdrr.io/rforge/climtrends/man/findOutliers.Tietjen.Moore.test.html

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1-21.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16): Asq Press.

Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems, 10*(2), 231-238.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766.

Maddala, G. S. (1992). *Introduction to Econometrics* (2nd ed.). New York: MacMillan.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology, 43*(4), 907-912.

Morgan, L. (2016, April 15). Data Outliers: 10 Ways To Prevent Big Data Damage. Retrieved from https://www.informationweek.com/big-data/big-data-analytics/data-outliers-10-ways-to-prevent-big-data-damage/d/d-id/1325130,

NIST/SEMATECH. (2003a, 2013). e-Handbook of Statistical Methods: Detection of outliers. Retrieved from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

NIST/SEMATECH. (2003b, 2013). e-Handbook of Statistical Methods: Dixon's Q-test: Detection of a single outlier. Retrieved from http://195.134.76.37/applets/AppletQtest/Appl_Qtest2.html

NIST/SEMATECH. (2003c, 2013). e-Handbook of Statistical Methods: Generalized ESD Test for Outliers. Retrieved from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

NIST/SEMATECH. (2003d, 2013). e-Handbook of Statistical Methods: Tietjen-Moore Test for Outliers. Retrieved from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h2.htm

RAY, S. (2016, January 10). A Comprehensive Guide to Data Exploration. Retrieved from https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics, 25*(2), 165-172.

Soetewey, A. (2020). *Outliers detection in R*.

Stefansky, W. (1972). Rejecting outliers in factorial designs. *Technometrics, 14*(2), 469-479.

Tietjen, G. L., & Moore, R. H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics, 14*(3), 583-597.

Univariate and Multivariate Outliers. (2020, June 24). Retrieved from https://www.statisticssolutions.com/univariate-and-multivariate-outliers/