บทความวิจัย

ช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์สำหรับพารามิเตอร์ของการแจกแจงปัวซงในการแจกแจงปัวซงค่าศูนย์เฟ้อ

พัทธ์ชนก ศรีสุรเดชชัย* และ กฤตนัน ตันประสงค์รัตน์
ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0 2564 4444 ต่อ 2101 กด 106 อีเมล: patchanok@mathstat.sci.tu.ac.th DOI: 10.14416/j.kmutnb.2020.12.013
รับเมื่อ 21 กรกฎาคม 2563 แก้ไขเมื่อ 1 ตุลาคม 2563 ตอบรับเมื่อ 5 พฤศจิกายน 2563 เผยแพร่ออนไลน์ 23 ธันวาคม 2563

บทคัดย่อ

ข้อมูลจำนวนนับถูกพบได้ทั่วไปในหลายสถานการณ์ และมักนิยมใช้การแจกแจงปัวซงในการอธิบายการเกิดเหตุการณ์ที่สนใจ แต่ในบางเหตุการณ์ ค่าสังเกตศูนย์เกิดขึ้นเกินกว่าที่จะถูกพิจารณาว่ามีการแจกแจงปัวซงตามปกติได้ หนึ่งในการแจกแจงความน่าจะเป็นที่นิยมมากที่สุดที่ถูกประยุกต์ใช้กับข้อมูลที่มีลักษณะดังกล่าว คือ การแจกแจงปัวซงค่าศูนย์เฟ้อ (ZIP) ในการทบทวนวรรณกรรม งานวิจัยส่วนใหญ่เน้นไปที่พารามิเตอร์ของการแจกแจงแบร์นูลลีซึ่งเป็นส่วนประกอบหนึ่งของ ZIP ในงานวิจัยนี้จึงได้เสนอ การประมาณค่าแบบช่วงของพารามิเตอร์ของการแจกแจงปัวซงใน ZIP เมื่อพารามิเตอร์ของการแจกแจงแบร์นูลลีไม่ทราบค่า โดยใช้ฟังก์ชันภาวะน่าจะเป็นโพรไฟล์กำจัดพารามิเตอร์รบกวน โดยการศึกษาจะแบ่งออกเป็น การพิสูจน์ทางคณิตศาสตร์ และการศึกษาเชิงจำลอง การวัดประสิทธิภาพของช่วงจะพิจารณาจากค่าความน่าจะเป็นคุ้มรวม (CP) และความยาวเฉลี่ย (AL) ของช่วงโดยวิธีมอนติคาร์โล ผลการศึกษาพบว่า โดยภาพรวม ช่วงที่เสนอขึ้นให้ค่า CP ใกล้เคียงกับสัมประสิทธิ์ความเชื่อมั่นที่ต้องการ และเมื่อค่าที่แท้จริงของพารามิเตอร์ของการแจกแจงปัวซงมีค่าเพิ่มมากขึ้น ช่วงที่นำเสนอมีประสิทธิภาพดีถึงแม้ตัวอย่างมีขนาดเล็ก

คำสำคัญ: ข้อมูลจำนวนนับ การแจกแจงปัวซง การจำลองมอนติคาร์โล ช่วงความเชื่อมั่น ภาวะน่าจะเป็นโพรไฟล์

การอ้างอิงบทความ: พัทธ์ชนก ศรีสุรเดชชัย และ กฤตนัน ตันประสงค์รัตน์, “ช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์สำหรับพารามิเตอร์ของการแจกแจงปัวซงในการแจกแจงปัวซงค่าศูนย์เฟ้อ,” *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 31, ฉบับที่ 2, หน้า 319–331, เม.ย.–มิ.ย. 2564.

Research Article

Profile-likelihood Based Confidence Intervals for the Poisson Parameter of Zero-inflated Poisson Distribution

Patchanok Srisuradetchai* and Kittanan Tonprasongrat
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Rangsit Campus, Pathum-Thani, Thailand

* Corresponding Author, Tel. 0 2564 4444 Ext. 2101 Press 106, E–mail: patchanok@mathstat.sci.tu.ac.th DOI: 10.14416/j.kmutnb.2020.12.013
Received 21 July 2020; Revised 1 October 2020; Accepted 5 November 2020; Published online: 23 December 2020

Abstract

Count data are commonly encountered in various real-life situations and researchers usually employ the Poisson distribution to elucidate such interesting events. Nevertheless, some occurrences possess too zeros to be reflected as a regular Poisson distribution. One of the most widely used probability distributions has been applied to data with excessive zeros is the Zero-inflated Poisson distribution (ZIP). In literature reviews, many devoted studies to the Bernoulli parameter, one component of the ZIP, and thus in this paper an interval estimation is proposed for the Poisson parameter in ZIP when the Bernoulli parameter is assumed to be unknown. The nuisance parameter is eliminated by a profile likelihood approach. The studies consist of mathematical proofs and simulations. The performance of proposed intervals is evaluated via the Coverage Probability (CP) and Average Length (AL) of confidence intervals, which were estimated by Monte-carlo methods. The results reveal that overall, the proposed interval produces the CP close to the desirable confidence coefficient. When the parameter of Poisson distribution becomes larger, the performance of profile likelihood-based confidence intervals is satisfied even though the sample is small.

Keywords: Count Data, Poisson Distribution, Monte-carlo Simulation, Confidence Interval, Profile Likelihood

*Please cite this article as: P. Srisuradetchai and K. Tonprasongrat, "Profile-likelihood based confidence intervals for the Poisson parameter of zero-inflated Poisson distribution," *The Journal of KMUTNB*, vol. 31, no. 2, pp. 319–331, Apr.–Jun. 2021 (in Thai).

1. บทนำ

การแจกแจงปัวซงเป็นการแจกแจงความน่าจะเป็นของตัวแปรสุ่มชนิดไม่ต่อเนื่อง (Discrete Random Variable) สำหรับข้อมูลจำนวนนับ (Count Data) ที่หมายถึง จำนวนครั้งที่เกิดเหตุการณ์ที่สนใจเกิดบนช่วงเวลาหนึ่งๆ หรือในพื้นที่หนึ่งๆ เช่น จำนวนคนไข้ที่เข้ารับการรักษาในโรงพยาบาลแห่งหนึ่งในช่วงเวลาหนึ่ง ให้ตัวแปรสุ่ม X แทน จำนวนครั้งของความสำเร็จที่การเกิดเหตุการณ์ที่สนใจซึ่งจะมีค่าเป็น 0, 1, 2, ... ที่มีพารามิเตอร์เป็น λ ซึ่งหมายถึง ค่าเฉลี่ยของจำนวนครั้งที่เกิดเหตุการณ์ที่สนใจ และสามารถเขียนแทนด้วยสัญลักษณ์ $X \sim \mathrm{Poi}(\lambda)$ สำหรับฟังก์ชันมวลความน่าจะเป็นของการแจกแจงปัวซง (พีเอ็มเอฟ) คือ $f(x,\lambda) = e^{-\lambda}\lambda^x/x!, x = 0, 1, \ldots$ ที่มีค่าคาดหมาย $E(X)$ และความแปรปรวน $Var(X)$ ของตัวแปรสุ่มเท่ากันเท่ากับ λ อย่างไรก็ตาม ในการประยุกต์ใช้กับข้อมูลจริงมักเกิดปัญหาความแปรปรวนมีค่ามากกว่าค่าคาดหมาย หรือที่เรียกว่าการกระจายเกินเกณฑ์ (Overdispersion) นอกจากนี้ ข้อมูลที่มีค่าศูนย์เป็นจำนวนมากกว่าปกติก็เป็นสาเหตุอย่างหนึ่งที่ทำให้เกิดภาวะการกระจายเกินเกณฑ์ได้ เช่น ข้อมูลจำนวนผู้ป่วยที่มีโรคแทรกซ้อนในเด็กอายุ 0–12 ปี ที่เป็นโรคปอดบวมในประเทศมาเลเซีย จำนวน 1,252 ราย โดยข้อมูลชุดนี้มีค่าสังเกตที่มีค่าเป็นศูนย์ถึง 62.3% [1] เนื่องจากมีค่าศูนย์มากกว่าปกติ ข้อมูลลักษณะนี้นิยมเรียกว่า มีค่าศูนย์เฟ้อ (Zero-inflated) ซึ่งทำให้ค่าคาดหมายและค่าความแปรปรวนมีค่าไม่เท่ากันจึงผิดตามข้อกำหนดของการแจกแจงปัวซง ในอดีต ค.ศ. 1992 Lambert [2] เป็นคนแรกที่ได้เสนอการแจงแจงปัวซงในกรณีพิเศษที่มีค่าสังเกตมีค่าเป็นศูนย์มากกว่าปกติ ซึ่งถูกเรียกว่า “การแจกแจงปัวซงค่าศูนย์เฟ้อ”

การแจกแจงปัวซงค่าศูนย์เฟ้อ (Zero-inflated Poisson Distribution; ZIP) เป็นการแจกแจงความน่าจะเป็นของตัวแปรสุ่มชนิดไม่ต่อเนื่องที่มี 2 พารามิเตอร์ คือ พารามิเตอร์ของการแจกแจงปัวซงหรือ λ และพารามิเตอร์ของการแจกแจงแบร์นูลลีหรือ π ซึ่งมีพีเอ็มเอฟเป็นดังสมการที่ (1)

$$f(x;\lambda,\pi) = [\pi + (1-\pi)e^{-\lambda}]^{I_{\{0\}}(x)}[(1-\pi)e^{-\lambda}\lambda^x/x!]^{1-I_{\{0\}}(x)} \quad (1)$$

โดยที่ $\lambda > 0$, $0 < \pi < 1$ และฟังก์ชัน $I_{\{0\}}(x)$ มีค่าเท่ากับ 1 เมื่อ $x = 0$ และมีค่าเท่ากับ 0 เมื่อ x มีค่าอื่นๆ โดยเขียนแทนด้วยสัญลักษณ์ $X \sim \mathrm{ZIP}(\lambda, \pi)$ สำหรับค่าคาดหมายและความแปรปรวนของ ZIP คือ $(1-\pi)\lambda$ และ $(1-\pi)\lambda(1-\pi\lambda)$ ตามลำดับ การแจกแจงนี้เป็นที่นิยมอย่างมากจึงได้ถูกนำไปประยุกต์ในหลากหลายสาขาวิชา เช่น ด้านคณิตศาสตร์ประกันภัย Boucher และคณะ [3] ใช้ ZIP เป็นตัวแบบการเรียกร้องค่าสินไหมของผู้ทำประกันภัย ในด้านการแพทย์ Böhning และคณะ [4] ได้ศึกษาข้อมูลดัชนี DMFT (Decayed Missing and Dilled Teeth) ในทางทันตกรรม

Beckett และคณะ [5] ได้เสนอการประมาณพารามิเตอร์แบบจุดของ ZIP ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) ได้ 2 สมการที่ใช้ในการหาตัวประมาณเป็น $\sum_{i=1}^{n} x_i[1-e^{-\hat{\lambda}}] = \hat{\lambda}(n-n_0)$ และ $\hat{\pi}_{ML} = (n_0 - ne^{-\hat{\lambda}})/(n - ne^{-\hat{\lambda}})$ ซึ่งไม่สามารถหารูปแบบปิด (Closed Form) สำหรับตัวประมาณของพารามิเตอร์ λ ได้ นอกจากนี้ Beckett ยังได้เสนอตัวประมาณที่ใช้วิธีโมเมนต์ (Method of Moment) ที่มีสูตรเป็น $\hat{\lambda}_{MM} = \bar{X} + (S^2/\bar{X}) - 1$ และ $\hat{\pi}_{MM} = (S^2 - \bar{X})/[\bar{X}^2 + S^2 - \bar{X}]$ โดยที่ $\bar{X}$ และ S^2 คือ ค่าเฉลี่ยและความแปรปรวนของตัวอย่าง Wagh และ Kamalja [6] ได้ประมาณพารามิเตอร์แบบจุดของ π ด้วยวิธี Probability Estimation (PE) ได้ผลลัพธ์เป็น $\hat{\pi} = (\hat{n}_0 - e^{-\hat{\lambda}_{MM}})/(1 - e^{-\hat{\lambda}_{MM}})$ โดยที่ $\hat{n}_0$ คือ สัดส่วนของจำนวนค่าสังเกตที่มีค่าเป็นศูนย์ต่อจำนวนค่าสังเกตทั้งหมด และ $\hat{\lambda}_{MM}$ คือ ค่าประมาณแบบจุดของพารามิเตอร์ λ ด้วยวิธีโมเมนต์

Xie และคณะ [7] ได้เสนอและเปรียบเทียบประสิทธิภาพของการทดสอบสมมติฐานทางสถิติ $H_0 : \pi = 0$ (ข้อมูลมาจากการแจกแจงปัวซง) และ $H_1 : \pi \neq 0$ (ข้อมูลมาจาก ZIP) รวมทั้งสิ้น 6 วิธี ได้แก่ การทดสอบสกอร์ [8] การทดสอบอัตราส่วนภาวะน่าจะเป็น [9] การทดสอบไคกำลังสอง การทดสอบโดยใช้ช่วงความเชื่อมั่น การทดสอบค็อกแครน (Cochran Test) [10] และการทดสอบ Rao-chakravarti [11] โดยวิธีมอนติคาร์โลที่ทำซ้ำจำนวน 1,000 รอบ โดยขนาดตัวอย่างที่ศึกษาเท่ากับ 10, 20 และ 50 ได้ผลลัพธ์ว่าการทดสอบที่ใช้ช่วงความเชื่อมั่นมีกำลังของการทดสอบ

(Power of the Test) ต่ำกว่าการทดสอบอื่นๆ ในขณะที่การทดสอบอื่นๆ มีประสิทธิภาพใกล้เคียงกัน Numna [12] ได้เสนอการทดสอบสมมติฐานทางสถิติ $H_0 : \pi = 0$ โดยใช้การทดสอบวัลด์ (Wald's Test) และได้นำมาเปรียบเทียบประสิทธิภาพกับวิธีที่เคยถูกเสนอก่อนหน้าพบว่า การทดสอบวัลด์มีประสิทธิภาพในการทดสอบใกล้เคียงกับการทดสอบค็อกแครน Paneru และคณะ [13] ได้ประมาณค่าแบบช่วงของค่าคาดหมายของ ZIP โดยใช้วิธีบูตสแทร็ป (Bootstrap) Thongchomphu และ Mayureesawan [14] ได้เสนอการประมาณค่าพารามิเตอร์แบบช่วงของสัมประสิทธิ์การแปรผัน (Coefficient of Variation; CV) ซึ่งพัฒนามาจากช่วงความเชื่อมั่นแบบเชิงเส้นกำกับของพารามิเตอร์ π Srisuradetchai และ Junnamtuam [15] ได้เปรียบเทียบช่วงความเชื่อมั่นแบบวัลด์ของพารามิเตอร์ π ในตัวแบบ ZIP และ ZAP (Zero-altered Poisson) ที่มีฟังก์ชันเชื่อมโยงที่แตกต่างกัน คือ ฟังก์ชันเชื่อมโยงลอจิต (Logit Link) ฟังก์ชันเชื่อมโยงโพรบิต (Probit Link) และฟังก์ชันเชื่อมโยงคอมพลีเมนทารีล็อก-ล็อก (Cloglog Link)

จากการทบทวนวรรณกรรมข้างต้นจะเห็นว่า งานวิจัยที่ผ่านมามุ่งเน้นให้ความสนใจพารามิเตอร์ π แต่งานวิจัยที่เกี่ยวข้องกับช่วงความเชื่อมั่นของ λ ซึ่งเป็นพารามิเตอร์หนึ่งของ ZIP ยังไม่มีผู้ศึกษา ในงานวิจัยนี้จึงสนใจที่ใช้แนวคิดการอนุมานทางสถิติที่ใช้ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) เข้ามาช่วยในการหาช่วงความเชื่อมั่นของ λ เมื่อ π ไม่ทราบค่า

แนวคิดของการอนุมานเชิงสถิติตามแนวคิดของ Fisher [16] นั้น การอนุมานเชิงสถิตินั้นจะขึ้นกับฟังก์ชันภาวะน่าจะเป็นเพียงอย่างเดียว สมมติให้ $X_1, X_2, X_3, \ldots, X_n$ เป็นตัวอย่างสุ่มจากการแจกแจงหนึ่งที่มีพารามิเตอร์ θ แล้วฟังก์ชันภาวะน่าจะเป็นเขียนแทนด้วย $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ โดยที่ตัวประมาณแบบภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ θ คือ $\hat{\theta}_{ML}$ ที่ทำให้ $L(\theta)$ มีค่าสูงที่สุดหรือ $\hat{\theta}_{ML} = \arg\max L(\theta)$ และมีฟังก์ชันอัตราส่วนภาวะน่าจะเป็น (Likelihood Ratio) เป็น $\tilde{L}(\theta) = L(\theta)/L(\hat{\theta}_{ML})$ ซึ่งจะทำให้ $0 \le \tilde{L}(\theta) \le 1$ และ $\tilde{L}(\hat{\theta}_{ML}) = 1$ นอกจากนี้ ยังสามารถเขียนในรูปของล็อกฟังก์ชันอัตราส่วนภาวะน่าจะเป็น ได้เป็น $\tilde{l}(\theta) = \log \tilde{L}(\theta) = \log L(\theta) - \log L(\hat{\theta}_{ML})$ ซึ่งจะทำให้ $-\infty \le \tilde{l}(\theta) \le 0$ และ $\tilde{l}(\hat{\theta}_{ML}) = 0$ ฟังก์ชันอัตราส่วนภาวะน่าจะเป็นนี้ถูกนำมาใช้สร้างช่วงความเชื่อมั่นของ θ โดย Fisher [16] ซึ่งมีนิยาม ดังนี้

$$\left\{\theta \middle| \tilde{L}(\theta) > c\right\} = \left\{\theta \middle| \frac{L(\theta)}{L(\hat{\theta}_{ML})} > c\right\} \quad (2)$$

โดยที่ค่าคงที่ c เป็นค่าที่สามารถเลือกได้ ส่วนมากจะอาศัยการแจกแจงเชิงเส้นกำกับ (Asymptotic Distribution) ของสถิติ Wilk [17] ในการกำหนด สถิติ Wilk นิยามดังนี้

$$W = -2\log \tilde{L}(\theta) = -2\tilde{l}(\theta) \quad (3)$$

ตัวแปรสุ่ม W ในสมการที่ (3) จะลู่เข้าสู่การแจกแจงเชิงเส้นกำกับ คือ การแจกแจงไคกำลังสองที่มีองศาเสรีเท่ากับ 1 ดังนั้น หาก c มีค่าเท่ากับ $\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)$ และเมื่อแทนลงในสมการที่ (2) จะได้ช่วงความเชื่อมั่นแบบสถิติอัตราส่วนภาวะน่าจะเป็นที่มีระดับความเชื่อมั่น $(1-\alpha)100\%$ ดังแสดงในสมการที่ (4)

$$\left\{\theta \middle| \tilde{L}(\theta) > \exp\left(-\chi^2_{1,(1-\alpha)}/2\right)\right\} \quad (4)$$

สำหรับการแจกแจง ZIP ที่ปรากฏในสมการที่ (1) มีพารามิเตอร์ 2 ตัว ในขณะที่งานวิจัยนี้สนใจ λ มีพารามิเตอร์ที่ไม่สนใจ คือ π วิธีหนึ่งที่นิยมในการกำจัดพารามิเตอร์ที่ไม่สนใจหรือพารามิเตอร์รบกวน (Nuisance Parameter) คือการใช้ภาวะน่าจะเป็นโพรไฟล์ (Profile Likelihood) ในที่นี้สมมติสุ่มตัวอย่างขนาด n จาก ZIP สามารถเขียนล็อกของฟังก์ชันภาวะน่าจะเป็นร่วมได้ดังนี้

$$\begin{aligned}\log L(\lambda,\pi;x_1,\ldots,x_n) = \\ n_0\log\left[\pi + (1-\pi)e^{-\lambda}\right] + (n-n_0)\log(1-\pi) \\ -\lambda(n-n_0) + \sum_{i=1,x_i\neq 0}^{n} x_i \log\lambda - \log \prod_{i=1,x_i\neq 0}^{n} x_i!\end{aligned} \quad (5)$$

โดยที่ n_0 แทน จำนวนค่าสังเกตที่เท่ากับ 0 และเพื่อกำจัด

พารามิเตอร์ π จะต้องหาตัวประมาณแบบภาวะน่าจะเป็นสูงสุดของ π นี้ โดยที่กำหนดให้ λ เป็นค่าคงที่ สมมติว่าได้ออกมาเป็น $\tilde{\pi}$ ซึ่ง $\tilde{\pi}$ จะมีเทอมของอีกพารามิเตอร์หนึ่ง (λ) ติดอยู่ หลังจากนั้นแทนค่า π ในสมการที่ (5) ด้วย $\tilde{\pi}$ แล้วจะได้

$$l_p(\lambda,\tilde{\pi}) = \log L_p(\lambda,\tilde{\pi}) \tag{6}$$

จะเห็นว่า สัญลักษณ์ $l_p(\lambda,\tilde{\pi})$ แทน ล็อกของฟังก์ชันภาวะน่าจะเป็นโพรไฟล์ดังแสดงในสมการที่ (6) และ $\tilde{\pi}$ นี้ต่างจากตัวประมาณแบบภาวะน่าจะเป็นสูงสุด $\hat{\pi}_{ML}$ ซึ่งจะไม่ติดเทอมของ λ

ฟังก์ชันอัตราส่วนภาวะน่าจะเป็นโพรไฟล์ (Profile Likelihood Ratio) คือ อัตราส่วนของฟังก์ชันภาวะน่าจะเป็นโพรไฟล์ต่อฟังก์ชันภาวะน่าจะเป็นโพรไฟล์ที่แทนค่าพารามิเตอร์ที่สนใจด้วยค่าประมาณภาวะน่าจะเป็นสูงสุดของ λ ด้วย $\hat{\lambda}_P = \arg\max l_p(\lambda,\tilde{\pi})$ กล่าวคือ

$$\tilde{L}_p(\lambda) = \frac{L_p(\lambda,\tilde{\pi})}{\max L_p(\lambda,\tilde{\pi})} = \frac{L_p(\lambda,\tilde{\pi})}{L_p(\hat{\lambda}_P,\tilde{\pi})} \tag{7}$$

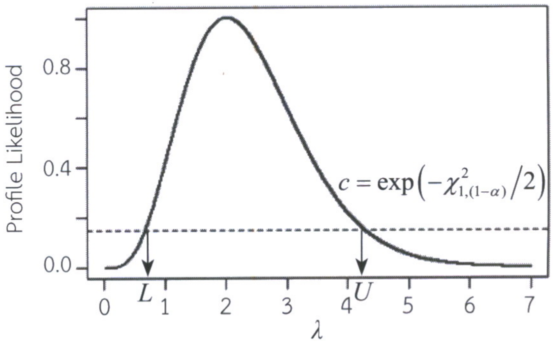
และมีล็อกของสมการที่ (7) เป็น $\tilde{l}_p(\lambda) = \log \tilde{L}_p(\lambda)$ โดยที่ $0 \le \tilde{L}_p(\lambda) \le 1$ และ $-\infty \le \tilde{l}_p(\lambda) \le 0$ เนื่องจากฟังก์ชันภาวะน่าจะเป็นโพรไฟล์ถูกมองว่าเป็นฟังก์ชันภาวะน่าจะเป็นอย่างหนึ่ง [18] จึงสามารถนำไปใช้สร้างช่วงได้ตามสมการที่ (4)

จากแนวความคิดทั้งหมดที่กล่าวตามข้างต้น งานวิจัยนี้จึงมีวัตถุประสงค์หลักในการหาช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์ของ λ ในการแจกแจง ZIP ที่ไม่ทราบค่าของ π โดยจะแบ่งการศึกษาออกเป็น 2 ส่วน คือ ส่วนของทฤษฎีที่พิสูจน์ด้วยหลักการทางคณิตศาสตร์และส่วนของการศึกษาเชิงจำลอง (Simulation Study)

2. วัสดุ อุปกรณ์และวิธีการวิจัย

เนื่องจากช่วงความเชื่อมั่น $(1-\alpha)\%$ แบบภาวะน่าจะเป็นโพรไฟล์เป็น

$$\left\{\lambda \middle| \tilde{L}_p(\lambda) \ge \exp\left(-\chi^2_{1,(1-\alpha)}/2\right)\right\} \tag{8}$$



รูปที่ 1 ขอบล่าง (L) และขอบบน (U) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ ที่ได้จาก $\tilde{L}_p(\lambda)$

ในทางทฤษฎีจะพิสูจน์ ดังต่อไปนี้

1) ตัวประมาณของ λ ที่ทำให้ $l_p(\lambda,\tilde{\pi})$ หรือสมการที่ (6) มีค่าสูงสุดหรือ $\hat{\lambda}_P = \arg\max l_p(\lambda,\tilde{\pi})$

2) กำหนด $g(\lambda) = \tilde{L}_p(\lambda)$ ขอบล่าง (Lower Limit) และขอบบน (Upper Limit) ของช่วงความเชื่อมั่น (L, U) หาได้จาก $L = g^{-1}\left(\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)\right)$ และ $U = g^{-1}\left(\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)\right)$ ตามลำดับ ทั้งนี้ $0 \le L < U < \infty$ ดังแสดงในรูปที่ 1

3) ขอบล่างและบนของช่วงความเชื่อมั่นจะหาได้ก็ต่อเมื่อ $\lim_{\lambda\to 0^+} \tilde{L}_p(\lambda)$ และ $\lim_{\lambda\to\infty} \tilde{L}_p(\lambda)$ เท่ากับศูนย์ เพราะค่าของฟังก์ชัน $g(\lambda)$ ต้องต่ำกว่าค่าคงที่ $c > 0$

และในการศึกษาเชิงจำลอง มีขั้นตอนดังนี้

1) จำลองประชากรขนาดใหญ่ $N = 10^7$ เสมือน ขนาดอนันต์ที่กำหนด λ เป็นค่าต่างๆ เป็น 1, 3, 5, 7, 9 และพารามิเตอร์ π เท่ากับ 0.1, 0.3, 0.5, 0.7, 0.9 จะได้กรณีที่เป็นไปได้ทั้งหมดของ (λ, π) เท่ากับ 25 แบบ โดยค่า π ที่กำหนดนั้นครอบคลุมเกือบทั้งปริภูมิพารามิเตอร์ (Parameter Space) สำหรับ λ ที่มีค่ามากกว่า 9 นั้น จะให้ผลที่มีความแตกต่างเพียงเล็กน้อยจากผลเมื่อ $\lambda = 5, 7, 9$ (ในภายหลังจะเห็นว่า ค่า $\lambda = 5, 7, 9$ ให้ผลลัพธ์ใกล้เคียงกันแล้ว)

2) ในแต่ละกรณีของประชากร จะสุ่มตัวอย่างขนาด n โดยที่ n เท่ากับ 10, 30, 50, 100 และ 200

3) จากตัวอย่างในแต่ละกรณีของ (λ, π, n) นำไปหาช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์ 95% กระทำซ้ำ

จำนวน 10,000 รอบ

4) จากขั้นตอนที่ 3 จะหาค่าประมาณของความน่าจะเป็นคุ้มรวม (Coverage Probability; CP) ได้จากสูตร

$$CP \approx \sum_{i=1}^{10,000} I_{[L_i,U_i]}(\lambda) \Big/ 10,000$$

โดยที่ $[L_i, U_i]$ แทนขอบล่างและบนของช่วงความเชื่อมั่นในรอบที่ i, i = 1, 2, ..., 10,000 และ $I_{[L_i, U_i]}(\lambda) = 1$ หาก λ ตกอยู่ในช่วง $[L_i, U_i]$ และเท่ากับศูนย์ในกรณีอื่นๆ สำหรับค่าประมาณความยาวช่วงโดยเฉลี่ย (Average Length; AL) คำนวณจาก

$$AL \approx \sum_{i=1}^{10,000} (U_i - L_i) \Big/ 10,000$$

3. ผลการทดลอง

สำหรับผลการศึกษาจะแบ่งเป็นเชิงทฤษฎีและเชิงการจำลอง สูตรที่ใช้ในการคำนวณช่วงความเชื่อมั่นแบบโพรไฟล์จะแสดงในบทตั้ง 2 โดยอาศัยผลของบทตั้ง 1 และได้กล่าวถึงทฤษฎีที่ใช้ในการตรวจสอบว่า ช่วงความเชื่อมั่นได้สามารถหาได้หรือไม่

3.1 ผลการศึกษาทางคณิตศาสตร์

ในส่วนนี้ สามารถแสดงได้เป็นทฤษฎีบทและบทตั้ง ดังนี้

บทตั้ง 1 กำหนดให้ $x_1, x_2, \ldots, x_n$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงปัวซงค่าศูนย์เฟ้อ $\mathrm{ZIP}(\lambda, \pi)$ โดยที่ไม่ทราบค่าพารามิเตอร์ λ และ π แล้วฟังก์ชันภาวะน่าจะเป็นโพรไฟล์ $L_p(\lambda, \tilde{\pi})$ ซึ่งมี $\tilde{\pi} = \dfrac{n_0 - ne^{-\lambda}}{n - ne^{-\lambda}}$ จะมีค่าสูงสุดเมื่อ λ เป็นรากของสมการ

$$(n-n_0)\lambda - \sum_{i=1, x_i \neq 0}^{n} x_i (1-e^{-\lambda}) = 0$$

พิสูจน์ จากสมการที่ (5) เมื่อกำหนดให้ λ เป็นค่าคงที่แล้วหาอนุพันธ์เทียบกับ π จะได้

$$\frac{\partial}{\partial \pi} \log L(\lambda, \pi) = n_0 \frac{\partial}{\partial \pi} \log\left[\pi + (1-\pi)e^{-\lambda}\right] + (n-n_0)\frac{\partial}{\partial \pi}\log(1-\pi)$$

ให้ $\dfrac{\partial}{\partial \pi} \log L(\lambda, \tilde{\pi}) = 0$ จะได้ $\tilde{\pi}$ ตามบทตั้ง และเมื่อแทน π ด้วย $\tilde{\pi}$ ลงใน $\log L(\lambda, \pi; x)$ จะได้

$$\begin{aligned}
\log L_P(\lambda, \tilde{\pi}; x) &= n_0 \log\left[\tilde{\pi} + (1-\tilde{\pi})e^{-\lambda}\right] + (n-n_0)\log(1-\tilde{\pi}) \\
&\quad -\lambda(n-n_0) + \sum_{i=1, x_i \neq 0}^{n} x_i \log\lambda - \log \prod_{i=1, x_i \neq 0}^{n} x_i! \\
&= n_0 \log\left[\left(\frac{n_0 - ne^{-\lambda}}{n - ne^{-\lambda}}\right) + \left(\frac{n-n_0}{n-ne^{-\lambda}}\right)e^{-\lambda}\right] + \\
&\quad (n-n_0)\log\left[\frac{n-n_0}{n-ne^{-\lambda}}\right] - \lambda(n-n_0) + \sum_{i=1, x_i \neq 0}^{n} x_i \log\lambda + c_1 \\
&= n_0 \log\left(\frac{n_0}{n}\right) + (n-n_0)\log\left(\frac{n-n_0}{n-ne^{-\lambda}}\right) - \\
&\quad \lambda(n-n_0) + \sum_{i=1, x_i \neq 0}^{n} x_i \log\lambda + c_2
\end{aligned}$$

โดยที่ c_1 และ c_2 เป็นเทอมที่ไม่ติด λ และเมื่อให้ $\dfrac{\partial}{\partial \lambda} \log L_P(\lambda, \tilde{\pi}) = 0$ จะได้

$$\begin{aligned}
\frac{-(n-n_0)(ne^{-\lambda})}{n-ne^{-\lambda}} - (n-n_0) + \frac{\sum_{i=1, x_i \neq 0}^{n} x_i}{\lambda} &= 0 \\
\lambda(n-n_0)e^{-\lambda} + (1-e^{-\lambda})\left(\lambda(n-n_0) - \sum_{i=1, x_i \neq 0}^{n} x_i\right) &= 0 \\
(n-n_0)\lambda - \sum_{i=1, x_i \neq 0}^{n} x_i(1-e^{-\lambda}) &= 0
\end{aligned}$$

สมการข้างต้นไม่สามารถเขียนในรูปปิดได้ การหารากของสมการที่ไม่ใช่เชิงเส้นนี้อาจทำได้โดยง่ายหากเรียกใช้ฟังก์ชัน Uniroot.all ในไลบรารี RootSolve [19], [20] หารากของสมการโดยวิธีของ Newton-raphson และให้รากของสมการแทนด้วยสัญลักษณ์ $\hat{\lambda}_p$

บทตั้ง 2 กำหนดให้ $x_1, x_2, \ldots, x_n$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงปัวซงค่าศูนย์เฟ้อ $\mathrm{ZIP}(\lambda, \pi)$ โดยที่ไม่ทราบค่าพารามิเตอร์ λ และ π ช่วงความเชื่อมั่น $(1-\alpha)\%$ แบบภาวะน่าจะเป็นโพรไฟล์ของ λ คือ เซตของค่า λ ที่สอดคล้องกับอสมการ

$$K_1 + K_2 \log\lambda + K_3\left[\lambda + \log(1-e^{-\lambda})\right] \geq 0$$

โดยที่

$$K_1 = (n-n_0)\left[\hat{\lambda}_P + \log(1-e^{-\hat{\lambda}_P})\right] - \sum_{i=1,x_i\neq 0}^{n} x_i \log\hat{\lambda}_P$$

$+\chi^2_{1,(1-\alpha)}/2,\ K_2 = \sum_{i=1,x_i\neq 0}^{n} x_i,\ K_3 = n_0 - n$ และมี $\hat{\lambda}_P$ จากบทตั้งที่ 1

พิสูจน์ จะหา $\tilde{L}_p(\lambda,\tilde{\pi})$ ก่อน ดังนี้

$$L(\lambda,\tilde{\pi}) = \left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n-n_0}{n-ne^{-\lambda}}\right)^{n-n_0} e^{-\lambda(n-n_0)}\frac{\lambda^{\sum_{i=1,x_i\neq 0}^{n} x_i}}{\prod_{i=1,x_i\neq 0}^{n} x_i!}$$

และ

$$L(\hat{\lambda}_P,\tilde{\pi}) = \left[\frac{n_0}{n}\right]^{n_0}\left[\frac{n-n_0}{n-ne^{-\hat{\lambda}_P}}\right]^{n-n_0} e^{-\hat{\lambda}_P(n-n_0)}\frac{\hat{\lambda}_P^{\sum_{i=1,x_i\neq 0}^{n} x_i}}{\prod_{i=1,x_i\neq 0}^{n} x_i!}$$

จะได้ว่า

$$\frac{L(\lambda,\tilde{\pi})}{L(\hat{\lambda}_P,\tilde{\pi})} = \left(\frac{n-ne^{-\hat{\lambda}_P}}{n-ne^{-\lambda}}\right)^{n-n_0} e^{-\lambda(n-n_0)+\hat{\lambda}_P(n-n_0)}\left(\frac{\lambda}{\hat{\lambda}_P}\right)^{\sum_{i=1,x_i\neq 0}^{n} x_i}$$

$$= \left(\frac{1-e^{-\hat{\lambda}_P}}{1-e^{-\lambda}}\right)^{n-n_0} e^{(n-n_0)(\hat{\lambda}_P-\lambda)}\left(\frac{\lambda}{\hat{\lambda}_P}\right)^{\sum_{i=1,x_i\neq 0}^{n} x_i}$$

และมี

$$\log\tilde{L}_p(\lambda) = \log\frac{L(\lambda,\tilde{\pi})}{L(\hat{\lambda}_P,\tilde{\pi})} = (n-n_0)\log\left(\frac{1-e^{-\hat{\lambda}_P}}{1-e^{-\lambda}}\right) +$$

$$(n-n_0)(\hat{\lambda}_P-\lambda) + \sum_{i=1,x_i\neq 0}^{n} x_i \log\left(\frac{\lambda}{\hat{\lambda}_P}\right)$$

และจากสมการที่ (8) หรือ $\log\tilde{L}_p(\lambda) \geq -\chi^2_{1,(1-\alpha)}/2$ เมื่อจัดรูปจะได้

$$K_1 + K_2\log\lambda + K_3\left[\lambda + \log(1-e^{-\lambda})\right] \geq 0$$

คำตอบอสมการข้างต้นนี้สามารถแก้ได้ไม่ยาก ในภาคผนวกได้แสดงฟังก์ชันในโปรแกรม R เพื่อหาคำตอบอสมการ (ช่วงความเชื่อมั่นแบบโพรไฟล์)

ทฤษฎีบท 1 กำหนดให้ $x_1, x_2, \ldots, x_n$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงปัวซงค่าศูนย์เฟ้อ ZIP(λ, π) โดยที่ไม่ทราบค่าพารามิเตอร์ λ และ π แล้ว

$$\lim_{\lambda\to 0^+}\tilde{L}_p(\lambda,\tilde{\pi}) = 1 \text{ และ } \lim_{\lambda\to\infty}\tilde{L}_p(\lambda,\tilde{\pi}) = 0$$

เมื่อ $\sum_{i=1}^{n} x_i = n - n_0$ และ

$$\lim_{\lambda\to 0^+}\tilde{L}_p(\lambda,\tilde{\pi}) = 0 \text{ และ } \lim_{\lambda\to\infty}\tilde{L}_p(\lambda,\tilde{\pi}) = 0$$

เมื่อ $\sum_{i=1}^{n} x_i > n - n_0$

พิสูจน์ พิจารณา

$$\lim_{\lambda\to 0^+}\frac{L(\lambda,\tilde{\pi})}{L(\hat{\lambda}_P,\tilde{\pi})} = \frac{e^{(n-n_0)\hat{\lambda}_P}(1-e^{-\hat{\lambda}_P})^{n-n_0}}{\hat{\lambda}_P^{\sum x_i}} \times \lim_{\lambda\to 0^+}\frac{1}{e^{(n-n_0)\lambda}}$$

$$\times \lim_{\lambda\to 0^+}\frac{\lambda^{\sum x_i}}{(1-e^{-\lambda})^{n-n_0}}$$

$$= \frac{e^{(n-n_0)\hat{\lambda}_P}(1-e^{-\hat{\lambda}_P})^{n-n_0}}{\hat{\lambda}_P^{\sum x_i}} \times 1 \times \lim_{\lambda\to 0^+}\frac{\lambda^{\sum x_i}}{(1-e^{-\lambda})^{n-n_0}}$$

พิจารณาเทอมสุดท้ายในกรณีที่ $\sum_{i=1}^{n} x_i = n - n_0$ จะได้ว่า

$$\lim_{\lambda\to 0^+}\left(\frac{\lambda}{1-e^{-\lambda}}\right)^{n-n_0} = \left(\lim_{\lambda\to 0^+}\frac{\lambda}{1-e^{-\lambda}}\right)^{n-n_0}$$

โดยกฎของโลปิตาล (L'Hôpital's Rule) จะได้ว่า

$$\left(\lim_{\lambda\to 0^+}\frac{1}{e^{-\lambda}}\right)^{n-n_0} = \left(\lim_{\lambda\to 0^+} e^{\lambda}\right)^{n-n_0} = 1$$

และในกรณีที่ $\sum x_i > n - n_0$ หากพิจารณาเทอม $\lim_{\lambda\to 0^+}\lambda^{\sum x_i}/(1-e^{-\lambda})^{n-n_0}$ จะเห็นว่าลิมิตอยู่ในรูปแบบยังไม่กำหนด (Indeterminate Form) ซึ่งเป็น 0/0 จึงใช้กฎของโลปิตาล ในที่นี้กำหนด $h(\lambda) = \lambda^{\sum x_i}$ เมื่อหาอนุพันธ์อันดับที่ 1, 2, 3, ..., $\left(\sum x_i\right)$ สามารถเขียนรูปทั่วไปเป็น

$$h^{(1)}(\lambda) = \left(\sum x_i\right)!/\left(\sum x_i - 1\right)!\lambda^{(\sum x_i - 1)} = A_1\lambda^{(\sum x_i - 1)}$$

$$h^{(2)}(\lambda) = \left(\sum x_i\right)!/\left(\sum x_i - 2\right)!\lambda^{(\sum x_i - 2)} = A_2\lambda^{(\sum x_i - 2)}$$

$$\vdots$$

$$h^{(k)}(\lambda) = \left(\sum x_i\right)!/\left(\sum x_i - k\right)!\lambda^{(\sum x_i - k)} = A_k\lambda^{(\sum x_i - k)}$$

$$\vdots$$

$$h^{(\sum x_i)}(\lambda) = \left(\sum x_i\right)!\lambda^{(\sum x_i - \sum x_i)} = A_{\sum x_i}\lambda^0$$

โดยที่ A_i, i = 1, 2, 3, ..., $\left(\sum x_i\right)$ เป็นค่าคงที่ที่ไม่ขึ้นกับ λ และในทำนองเดียวกัน กำหนด $g(\lambda)=(1-e^{-\lambda})^{n-n_0}$ เมื่อหาอนุพันธ์อันดับที่ k, k = 1, 2, 3, ..., $\left(\sum x_i\right)$ โดยที่ $\sum x_i > n-n_0$ จะได้รูปทั่วไปซึ่งยังติดในเทอมของ $e^{-\lambda}$ และ $1-e^{-\lambda}$ เท่านั้น โดยจะได้ $g^{(k)}(\lambda)$ =

$$\begin{cases}\sum_{i=1}^{k} B_i^{(k)} e^{-i\lambda}(1-e^{-\lambda})^{(n-n_0)-i}, k=1,2,\ ...,(n-n_0-1)\\ \sum_{i=1}^{(n-n_0)} C_i^{(k)} e^{-i\lambda}(1-e^{-\lambda})^{(n-n_0)-i}, k=(n-n_0),...,\sum x_i\end{cases}$$

โดยที่ B_i, i = 1, ..., k และ C_i, i = 1, ..., $n-n_0$ เป็นค่าคงที่ที่ไม่ขึ้นกับ λ จะสังเกตว่า เทอม

$$\sum_{i=1}^{(n-n_0)} C_i^{(k)} e^{-i\lambda}(1-e^{-\lambda})^{(n-n_0)-i}$$

ขึ้นอยู่กับค่า k อย่างแฝง กล่าวคือ ถึงแม้ว่าค่า k ต่างกัน เช่น k และ k' จำนวนพจน์ในผลรวมยังคงเท่ากันเท่ากับ $n-n_0$ เทอม แต่ $C_i^{(k)}$ และ $C_i^{(k')}$ ไม่เท่ากัน และเช่นเดียวกันสำหรับ $\sum_{i=1}^{k} B_i^{(k)} e^{-i\lambda}(1-e^{-\lambda})^{(n-n_0)-i}$ ที่มี $B_i^{(k)}$ และ $B_i^{(k')}$ มีค่าต่างกัน นอกจากนี้ หากแทน λ ใน $g^{(k)}(\lambda)$ ด้วยศูนย์จะได้

$$g^{(k)}(0)=\begin{cases}0 & , k=1,2,\ ...,(n-n_0-1)\\ C_{(n-n_0)}, & k=(n-n_0),...,\sum x_i\end{cases}$$

กล่าวคือ หากหาอนุพันธ์อันที่ $k = (n-n_0)$, ..., $\sum x_i$ จะมีแต่พจน์สุดท้ายของ $g^{(k)}(\lambda)$ ที่ไม่ติด $(1-e^{-\lambda})$ จึงทำให้ $g^{(k)}(0) \neq 0$ ดังนั้น

$$\lim_{\lambda\to 0^+}\frac{h(\lambda)}{g(\lambda)} = \lim_{\lambda\to 0^+}\frac{h^{(n-n_0)}(\lambda)}{g^{(n-n_0)}(\lambda)} = \lim_{\lambda\to 0^+}\frac{A_{(n-n_0)}\lambda^{\left(\sum x_i-(n-n_0)\right)}}{C_{(n-n_0)}}=0$$

และ

$$\lim_{\lambda\to\infty}\frac{L(\lambda,\tilde{\pi})}{L(\hat{\lambda}_P,\tilde{\pi})} = \frac{e^{(n-n_0)\hat{\lambda}_P}(1-e^{-\hat{\lambda}_P})^{n-n_0}}{\hat{\lambda}_P{}^{\sum x_i}}\times\lim_{\lambda\to\infty}\frac{1}{(1-e^{-\lambda})^{n-n_0}} \times\lim_{\lambda\to\infty}\frac{\lambda^{\sum x_i}}{e^{(n-n_0)\lambda}} = \frac{e^{(n-n_0)\hat{\lambda}_P}(1-e^{-\hat{\lambda}_P})^{n-n_0}}{\hat{\lambda}_P{}^{\sum x_i}}\times 1\times\lim_{\lambda\to\infty}\frac{\lambda^{\sum x_i}}{e^{(n-n_0)\lambda}}$$

พิจารณาเทอม $\lim_{\lambda\to\infty}\frac{\lambda^{\sum x_i}}{e^{(n-n_0)\lambda}} = \lim_{\lambda\to\infty}\frac{h(\lambda)}{e^{(n-n_0)\lambda}}$ จะเห็นว่าลิมิตอยู่ในรูปแบบยังไม่กำหนด (∞/∞) ให้ $r(\lambda) = e^{(n-n_0)\lambda}$ แล้วเมื่อหาอนุพันธ์อันดับที่ 1, 2, 3, ..., $\left(\sum x_i\right)$ จะสามารถเขียนในรูปทั่วไปได้เป็น

$$\begin{aligned} r^{(1)}(\lambda) &= (n-n_0)e^{(n-n_0)\lambda}\\ r^{(2)}(\lambda) &= (n-n_0)^2 e^{(n-n_0)\lambda}\\ &\vdots\\ r^{(k)}(\lambda) &= (n-n_0)^k e^{(n-n_0)\lambda}\\ &\vdots\\ r^{\left(\sum x_i\right)}(\lambda) &= (n-n_0)^{\left(\sum x_i\right)} e^{(n-n_0)\lambda}\end{aligned}$$

ดังนั้น

$$\lim_{\lambda\to\infty}\frac{h(\lambda)}{r(\lambda)} = \lim_{\lambda\to\infty}\frac{h^{\left(\sum x_i\right)}(\lambda)}{r^{\left(\sum x_i\right)}(\lambda)} = \lim_{\lambda\to\infty}\frac{\left(\sum x_i\right)!}{(n-n_0)^{\left(\sum x_i\right)}e^{(n-n_0)\lambda}}=0$$

จากการพิสูจน์ข้างต้น จะได้ว่าขอบล่างและบนของช่วงในกรณีที่ $\sum x_i > n-n_0$ จะสามารถหาค่าได้เสมอ เนื่องจาก $\lim_{\lambda\to 0^+}\tilde{L}_p(\lambda,\tilde{\pi})=0$ และ $\lim_{\lambda\to\infty}\tilde{L}_p(\lambda,\tilde{\pi})=0$ หรือมีค่า λ ที่ทำให้สมการที่ (8) เป็นจริง (พิจารณารูปที่ 1 ประกอบ) เพราะเส้นโค้ง $\tilde{L}_p(\lambda,\tilde{\pi})$ มีค่าที่ต่ำกว่า $\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)$ ในทั้งสองด้านสำหรับทุกค่าของ α ในขณะที่ $\sum x_i = n-n_0$ หรือข้อมูลที่มีเพียงค่าสังเกตศูนย์และหนึ่งเท่านั้น เช่น (1,1,1,0,0,0,0,0) กรณีนี้ $\sum x_i$ เท่ากับ $n-n_0$ ซึ่งเท่ากับ 3 ขอบบนของช่วงจะสามารถหาได้เสมอ แต่ขอบล่างจะถูกกำหนดให้เท่า 0 เนื่องจาก $\lim_{\lambda\to 0^+}\tilde{L}_p(\lambda,\tilde{\pi})=1$ (ไม่ได้ต่ำกว่า $\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)$) ดังนั้น ช่วงความเชื่อมั่นในกรณีนี้จะอยู่ในรูปของ $(0, U)$ โดยที่ U เป็นราก (เดียว) ของสมการ $\tilde{L}_p(\lambda)=\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)$

3.2 ผลการศึกษาเชิงการจำลอง

ค่าประมาณความน่าจะเป็นคุ้มรวม (CP) และค่าความยาวช่วงโดยเฉลี่ย (AL) ของช่วงความเชื่อมั่น 95% ของ λ แสดงดังในตารางที่ 1 โดยภาพรวมพบว่า พารามิเตอร์ของการแจกแจงปัวซง (λ) และของการแจกแจงแบร์นูลลี (π) ส่ง

ผลต่อทั้ง AL และ CP โดยเฉพาะเมื่อตัวอย่างมีขนาดเล็กหรือ $n = 10$ แต่เมื่อ $n \geq 30$ ค่าของ CP เข้าใกล้และมีค่ารอบๆ สัมประสิทธิ์ความเชื่อมั่น 0.95 หากพิจารณาค่า AL จะเห็นว่าลดลงอย่างชัดเจนเมื่อขนาดตัวอย่างเพิ่มขึ้น

ตารางที่ 1 ค่าประมาณความน่าจะเป็นคุ้มรวม (ความยาวช่วงโดยเฉลี่ย) ของช่วงความเชื่อมั่น 95% ของ λ

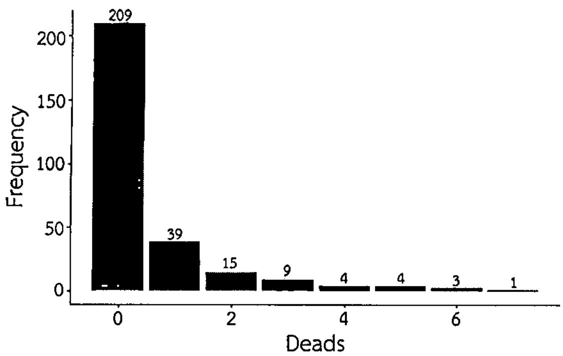
λ	π	n				
		10	30	50	100	200
1	0.1	0.9367 (1.9810)	0.9452 (1.1580)	0.9501 (0.9007)	0.9513 (0.6383)	0.9500 (0.4514)
	0.3	0.9363 (2.2581)	0.9449 (1.3139)	0.9461 (1.0234)	0.9488 (0.7231)	0.9469 (0.5119)
	0.5	0.9459 (2.6705)	0.9409 (1.5615)	0.9448 (1.2094)	0.9538 (0.8551)	0.9474 (0.6071)
	0.7	0.9694 (3.3131)	0.9389 (2.0482)	0.9362 (1.5606)	0.9472 (1.1123)	0.9489 (0.7839)
	0.9	0.9766 (4.0210)	0.9692 (3.3071)	0.9528 (2.7524)	0.9362 (1.9328)	0.9447 (1.3618)
3	0.1	0.9521 (2.4829)	0.9532 (1.4254)	0.9489 (1.1036)	0.9525 (0.7793)	0.9512 (0.5514)
	0.3	0.9480 (2.8494)	0.9490 (1.6225)	0.9529 (1.2501)	0.9525 (0.8839)	0.9502 (0.6255)
	0.5	0.9477 (3.4413)	0.9503 (1.9348)	0.9439 (1.4879)	0.9486 (1.0499)	0.9482 (0.7407)
	0.7	0.9554 (4.4939)	0.9527 (2.5335)	0.9462 (1.9466)	0.9526 (1.3587)	0.9478 (0.9596)
	0.9	0.9758 (5.9990)	0.9512 (4.4954)	0.9448 (3.5245)	0.9538 (2.4459)	0.9533 (1.6788)
5	0.1	0.9466 (2.9889)	0.9527 (1.7202)	0.9488 (1.3314)	0.9513 (0.9412)	0.9483 (0.6647)
	0.3	0.9526 (3.4352)	0.9498 (1.9609)	0.9506 (1.5145)	0.9497 (1.0672)	0.9562 (0.7541)
	0.5	0.9465 (4.1777)	0.9494 (2.3335)	0.9464 (1.7964)	0.9487 (1.2674)	0.9490 (0.8957)
	0.7	0.9487 (5.5346)	0.9498 (3.0802)	0.9495 (2.3467)	0.9496 (1.6436)	0.9478 (1.1586)
	0.9	0.9378 (7.6464)	0.9477 (5.6078)	0.9494 (4.3211)	0.9492 (2.9202)	0.9499 (2.0309)

ตารางที่ 1 ค่าประมาณความน่าจะเป็นคุ้มรวม (ความยาวช่วงโดยเฉลี่ย) ของช่วงความเชื่อมั่น 95% ของ λ (ต่อ)

λ	π	n				
		10	30	50	100	200
7	0.1	0.9512 (3.4857)	0.9525 (2.0055)	0.9473 (1.5526)	0.9488 (1.0974)	0.9517 (0.7759)
	0.3	0.9459 (4.0044)	0.9456 (2.2858)	0.9498 (1.7664)	0.9522 (1.2463)	0.9508 (0.8796)
	0.5	0.9466 (4.8846)	0.9506 (2.7243)	0.9524 (2.0980)	0.9494 (1.4769)	0.9540 (1.0423)
	0.7	0.9470 (6.5077)	0.9500 (3.5889)	0.9489 (2.7394)	0.9505 (1.9148)	0.9520 (1.3461)
	0.9	0.9451 (8.9833)	0.9489 (6.9672)	0.9465 (5.0270)	0.9484 (3.4160)	0.9519 (2.3609)
9	0.1	0.9474 (3.9414)	0.9498 (2.2675)	0.9519 (1.7575)	0.9483 (1.2409)	0.9525 (0.8773)
	0.3	0.9497 (4.5200)	0.9495 (2.5838)	0.9474 (1.9970)	0.9518 (1.4097)	0.9491 (0.9953)
	0.5	0.9534 (5.5016)	0.9472 (3.0753)	0.9476 (2.3709)	0.9493 (1.6721)	0.9488 (1.1785)
	0.7	0.9515 (7.3131)	0.9519 (4.0423)	0.9490 (3.0967)	0.9503 (2.1707)	0.9542 (1.5276)
	0.9	0.9541 (10.1709)	0.9466 (7.3892)	0.9493 (5.6390)	0.9489 (3.8761)	0.9487 (2.6768)

หมายเหตุ: ค่า AL อยู่ในวงเล็บ

ค่าพารามิเตอร์ λ นี้จะมีผลต่อค่า CP ในลักษณะที่ว่าเมื่อ λ มีค่าน้อยหรือ $\lambda = 1$ ค่าของ CP จะขึ้นอยู่กับพารามิเตอร์ π ซึ่งหากมีค่าน้อยมาก ($\pi = 0.1$) หรือสูงมาก ($\pi = 0.9$) ค่าของ CP ก็มีแนวโน้มที่จะแตกต่างจาก 0.95 มากขึ้น เช่น กรณีที่ขนาดตัวอย่างเท่ากับ 10 และ $\pi = 0.1$ มีค่า CP เท่ากับ 0.9367 และค่า AL เท่ากับ 1.9810 แต่เมื่อ $\pi = 0.9$ มีค่า CP เท่ากับ 0.9766 และค่า AL เท่ากับ 4.0210 นอกจากนี้ จะเห็นว่า AL มีแนวโน้มเพิ่มขึ้นชัดเจนเมื่อ π มีค่าสูงขึ้น เช่น กรณีที่ π มีค่าน้อยกับ 0.1 และตัวอย่างขนาดเล็กเท่ากับ 10 เมื่อ $\lambda = 5$ ค่า AL เท่ากับ 2.9889 และเมื่อ $\lambda = 7$ ค่า AL เท่ากับ 3.4857 แต่เมื่อ λ มีค่าเพิ่มสูงขึ้น ค่าของ π จะส่งผลกระทบต่อ CP น้อยมากหรือไม่สามารถสังเกตเห็นได้ชัด เช่น กรณีที่ $\lambda = 9$ ในตัว

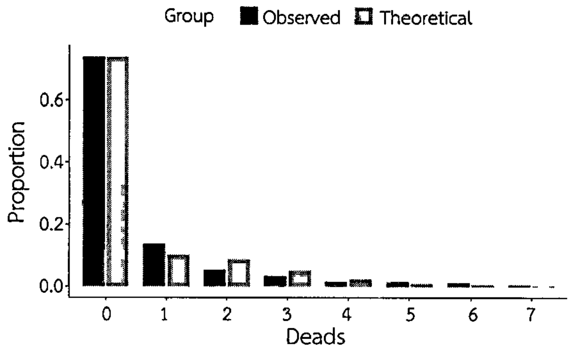


รูปที่ 2 ผู้เสียชีวิตจาก COVID-19 ในประเทศมาเลเซีย

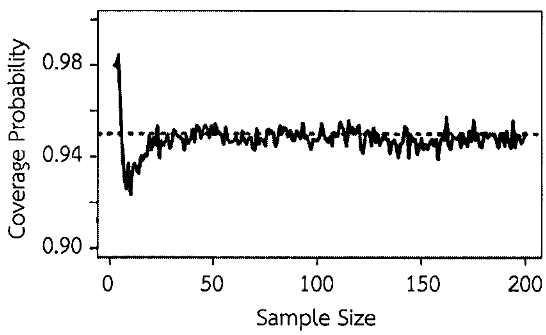
ตัวอย่างขนาดเล็กเท่ากับ 10 เมื่อ $\pi = 0.1$ ค่า CP เท่ากับ 0.9474 และเมื่อ $\pi = 0.9$ ค่า CP เท่ากับ 0.9541 ค่า CP ทั้งสองกรณีนี้ใกล้ 0.95 ในขณะที่ค่าพารามิเตอร์ π แตกต่างกันมาก

4. การประยุกต์ใช้กับข้อมูลจริง

จากข้อมูลองค์การอนามัยโลก (The World Health Organization) [21] จำนวนผู้เสียชีวิตจาก COVID-19 รายวัน ตั้งแต่วันที่ 3 มกราคม 2563 ถึง 12 ตุลาคม 2563 รวม 284 วัน ในประเทศมาเลเซีย แสดงในดังรูปที่ 2 จะเห็นว่าค่าสังเกตที่มีค่าเท่า 0 นั้น มีค่าสูงกว่าที่จะมีการแจกแจงปัวซงแบบธรรมดาได้ เมื่อหาค่าประมาณแบบภาวะน่าจะเป็นสูงสุด จะได้ว่า $\hat{\lambda}_{ML} = 1.7176$ และ $\hat{\pi}_{ML} = 0.6779$ หากนำไปประมาณค่าความน่าจะเป็นแล้วเปรียบเทียบกับสัดส่วนจากข้อมูลจริง ดังแสดงในรูปที่ 3 พบว่า สัดส่วนของจำนวนวันที่ไม่มีผู้เสียชีวิต (0.7359) ในข้อมูลจริงใกล้เคียงกับ $P(X = 0) = 0.678 + 0.322e^{-1.718} = 0.7359$ (ค่าประมาณมีความถูกต้องมากถึงทศนิยมตำแหน่งที่ 4) สำหรับค่าประมาณ $P(X=x), x=1,\ldots,7$ แตกต่างเล็กน้อยเมื่อเทียบกับข้อมูลจริง และเมื่อคำนวณช่วงความเชื่อมั่น 95% แบบภาวะน่าจะเป็นโพรไฟล์สำหรับ λ จะได้เป็น (1.3968, 2.076849) ซึ่งมีความยาวของช่วงเท่ากับ 0.68 ในกรณีนี้ หากทำการจำลองจำนวน 10,000 รอบ ขนาดตัวอย่างเท่ากับ 284 จาก $ZIP(\lambda = 1.717, \pi = 0.678)$ พบว่า CP เท่ากับ 0.9488 และ AL เท่ากับ 0.6828 ซึ่งใกล้เคียงกับ AL ของช่วงที่คำนวณจากข้อมูลจริงมาก จึงทำให้มั่นใจว่า ผลการศึกษาในตารางที่ 1 สามารถนำไปใช้ได้จริง



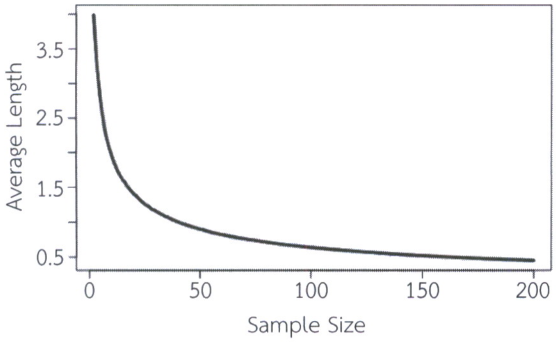
รูปที่ 3 เปรียบเทียบสัดส่วนผู้เสียชีวิตจาก COVID-19 ในประเทศมาเลเซียจากข้อมูลจริงและจากตัวแบบ $ZIP(\hat{\lambda}_{ML} = 1.7176, \hat{\pi}_{ML} = 0.6779)$



รูปที่ 4 ค่าประมาณความน่าจะเป็นคุ้มรวมของช่วงความเชื่อมั่น 95% ของ λ สำหรับ $ZIP(\lambda = 1, \pi = 0.1)$ เมื่อ $n = 2, 3, 4, \ldots, 200$

5. สรุป

การนำฟังก์ชันภาวะน่าจะเป็นโพรไฟล์มาใช้ในการหาช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นของ λ นั้น โดยภาพรวมช่วงที่ได้มีประสิทธิภาพดีเนื่องจากความน่าจะเป็นคุ้มรวม (CP) ที่ได้จากการศึกษาเชิงจำลองนั้นมีค่าไม่ห่างจาก 0.95 ค่า CP ที่น้อยสุดและมากที่สุดเป็น 0.9362 และ 0.9766 ตามลำดับ ซึ่งเกิดในกรณีที่ประชากรมีค่ามี λ ค่าน้อย ($\lambda = 1$) และเมื่อศึกษาเพิ่มเติมในกรณีนี้ซึ่งมีประชากรเป็น $ZIP(\lambda = 1, \pi = 0.1)$ และ n เท่ากับ 2, 3, ..., 200 ดังแสดงในรูปที่ 4 เมื่อ n เพิ่มขึ้นโดยประมาณเท่ากับ 25 ค่า CP จะมีค่าไม่แกว่งขึ้นลงห่างจาก 0.95 มากนัก และเป็นขนาดตัวอย่างที่ต่ำที่สุด



รูปที่ 5 ค่าความความยาวช่วงโดยเฉลี่ยของช่วงความเชื่อมั่น 95% ของ λ สำหรับ $ZIP(\lambda = 1, \pi = 0.1)$ เมื่อ $n = 2, 3, 4, ..., 200$

ที่หลังจากนี้ ค่า AL จะลดลงอย่างช้าๆ ดังแสดงในรูปที่ 5 หากขนาดตัวอย่างน้อยมากๆ ($n < 5$) ค่า CP มีค่าสูงเข้าใกล้ 1 และมีค่า AL สูงมากถึง 4 โดยประมาณ

สำหรับ ZIP ที่มีค่าพารามิเตอร์ $\lambda \geq 3$ ค่า CP จะเข้าใกล้ 0.95 ถึงแม้ตัวอย่างจะมีขนาดเล็กมาก ($n=10$) และแทบไม่ขึ้นกับค่าของ π ซึ่งเป็นพารามิเตอร์ที่แสดงค่าศูนย์เฟ้อ ดังนั้นหากในการวิเคราะห์ข้อมูลจริงพบว่า ค่าประมาณแบบจุดของ λ มีค่าสูง (มากกว่า 3) ผู้วิเคราะห์สามารถมั่นใจในช่วงความเชื่อมั่นที่นำเสนอนี้ได้

โดยสรุป จากการศึกษาเชิงคณิตศาสตร์พบว่า ช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์ที่นำเสนอสามารถหาทั้งขอบล่างและบนได้หาก $\sum x_i > n - n_0$ และในการศึกษาเชิงจำลองพบว่า ช่วงที่ได้มีประสิทธิภาพและสามารถนำไปใช้ได้จริงในหลายสถานการณ์ที่แม้ตัวอย่างจะมีขนาดเล็ก หรือพารามิเตอร์ของแบร์นูลลีมีค่าน้อย/มาก

ข้อเสนอแนะในการศึกษาต่อไป กรณีที่ข้อมูลจำนวนนับมีศูนย์เฟ้อและมีความสัมพันธ์กัน (Correlated Data) อาจเกิดขึ้นได้ในข้อมูลช่วงยาว (Longitudinal Data) ซึ่งมีความซับซ้อนในเชิงทฤษฎี ยังไม่มีงานวิจัยที่เกี่ยวข้องการประมาณค่าแบบช่วงและการทดสอบสมมติฐานสำหรับพารามิเตอร์ p และ λ ในบริบทของฟังก์ชันการแจกแจงความน่าจะเป็น จึงเป็นประเด็นหนึ่งที่น่าสนใจ สำหรับในบริบทของการวิเคราะห์การถดถอยสามารถอ่านได้จาก Zhang และคณะ [22]

เอกสารอ้างอิง

[1] W. M. A. W. Ahmad, S. A. Abdullah, K. Mokhtar, N. A. Aleng, N. Halim, and Z. Ali, “Application of zero inflated models for health sciences data,” *Journal of Advanced Scientific Research*, vol. 6, no. 2, pp. 39–44, 2015.

[2] D. Lambert, “Zero-inflated Poisson regression, with an application to defects in manufacturing,” *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[3] J. P. Boucher, M. Denuit, and M. Guillen, “Number of accidents or number of claim? An approach with zero-inflated Poisson models for panel data,” *The Journal of Risk and Insurance*, vol. 76, no. 4, pp. 821–846, 2009.

[4] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner, “The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology,” *Journal of the Royal Statistical Society*, vol. 162, no. 2, pp. 195–209, 1999.

[5] S. Beckett, J. Jee, T. Ncube, S. Pompilus, Q. Washington, A. Singh, and N. Pal, “Zero-inflated Poisson (ZIP) distribution: Parameter estimation and applications to model data from natural calamities,” *Involve a Journal of Mathematics*, vol. 7, no. 6, pp. 751–767, 2014.

[6] Y. S. Wagh and K. K. Kamalja, “Zero-inflated models and estimation in zero-inflated Poisson distribution,” *Communications in Statistics – Simulation and Computation*, vol. 47, no. 8, pp. 2248–2265, 2018.

[7] M. Xie, B. He, and T. N. Goh, “Zero-inflated Poisson model in statistical process control,” *Computational Statistics & Data Analysis*, vol. 38, no. 2, pp. 191–201, 2001.

[8] J. Vandenbroek, "A score test for zero inflation in a Poisson-distribution," *Biometrics*, vol. 51, no. 2, pp. 738–743, 1995.

[9] A. H. El-Shaarawi, "Some goodness-of-fit methods for the Poisson plus added zeros distribution," *Applied and Environmental Microbiology*, vol. 49, pp. 1304–1306, 1985.

[10] W. G. Cochran, "Some methods for strengthening the common tests," Biometrics, vol. 10, pp. 417–451, 1954.

[11] C. R. Rao and I. M. Chakravarti, "Some small sample tests of significance for a Poisson distribution," *Biometrics*, vol. 12, pp. 264–282, 1956.

[12] S. Numna, "Analysis of extra zero counts using zero-inflated Poisson models," M.S. thesis, Department Science in Mathematics and Statistics., Songkla University, Songkla, Thailand, 2009.

[13] K. Paneru, R. N. Padgett, and H. Chen, "Estimation of zero-inflated population mean: A bootstrapping approach," *Journal of Modern Applied Statistical Methods*, vol. 17, no. 1, 2018.

[14] P. Thongchomphu and T. Mayureesawan, "The confidence interval of the coefficient of variation for a zero-inflated Poisson, distribution," *The Journal of KMUTNB*, vol. 29, no. 4, pp. 652–666, 2019 (in Thai).

[15] P. Srisuradetchai and S. Junnumtuam, "Wald confidence intervals of the parameter in a bernoulli component of zero-inflated Poisson and zero-altered Poisson models with different link functions," *Science & Technology Asia (STA)*, vol. 25, no. 2, pp. 1–14, 2020 (in Thai).

[16] R. A. Fisher, *Statistical Methods and Scientific Inference*. New York: Macmillan, 1973.

[17] S. S. Wilk, "The large sample distribution of the likelihood ratio for testing composite hypotheses," *Annals Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[18] L. Held and D. S. Bové, *Applied Statistical Inference in Likelihood and Bayes*, 1st ed. London: Springer, 2014.

[19] K. Soetaert and P. M. Herman, *A Practical Guide to Ecological Modelling. Using R as a Simulation Platform*, Springer, 2009.

[20] K. Soetaert, *RootSolve: Nonlinear Root Finding, Equilibrium and Steady-state Analysis of Ordinary Differential Equations*, R package 1.6, 2009.

[21] World Health Organization. (2020, October 12). *WHO Coronavirus Disease (COVID-19) Dashboard* Available: https://covid19.who.int/

[22] W. Zhang, J. Wang, F Qian, and Y. Chen, "A joint mean-correlation modeling approach for longitudinal zero-inflated count data," *Brazilian Journal of Probability and Statistics*, vol. 34, no. 1, pp. 35–50, 2020.

ภาคผนวก โปรแกรม R

```
Like.Pro.CI <- function(dat){
  n0 <- sum(dat == 0)
  n <- length(dat)
  n1 <- n - n0
  dat.pos <- dat[which(dat > 0)]
  pro.llike <- function(lambda){
    pstr0 <- (n0 - n*exp(-lambda))/(n - n*exp(-lambda))
    n0*log(pstr0 + (1 - pstr0)*exp(-lambda)) +
       (n - n0)*log(1 - pstr0) - lambda*(n - n0)+
       log(lambda)*sum(dat.pos) - sum(log(factorial(dat.pos)))
  }
  solution <- maxLik(pro.llike, start = c(lambda = 1), method = "SANN")
  point <- solution$estimate
  K1 <- ((n-n0)*log(1-exp(-point))) + ((n-n0)*point) -
        (sum(dat.pos)*log(point)) + (qchisq(0.95, df = 1)/2)
  K2 <- sum(dat.pos)
  K3 <- (n0 - n)
  fun.1 <- function(x) K1 + K2*log(x) + K3*(log(1-exp(-x))+x)
  solution <- uniroot.all(fun.1, c(0.0001, 30), tol = 1e-10)
  return(solution)
}
```