บทความวิจัย

ช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์สำหรับพารามิเตอร์ของการแจกแจงเรขาคณิตในการแจกแจงเรขาคณิตค่าศูนย์เฟ้อ

พัทธ์ชนก ศรีสุรเดชชัย* และ กิตติมา แดงสุภา
ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0 2564 4444 ต่อ 2101 กด 106 อีเมล: patchanok@mathstat.sci.tu.ac.th DOI: 10.14416/j.kmutnb.2021.05.015
รับเมื่อ 17 กันยายน 2563 แก้ไขเมื่อ 18 ตุลาคม 2563 ตอบรับเมื่อ 6 พฤศจิกายน 2563 เผยแพร่ออนไลน์ 24 พฤษภาคม 2564

บทคัดย่อ

ในการประยุกต์ใช้เครื่องมือทางสถิติกับข้อมูลเชิงนับ บางครั้งค่าสังเกตศูนย์มีความถี่มากกว่าที่ควรจะเป็นสำหรับการแจกแจงที่ใช้ในการศึกษา การแจกแจงเรขาคณิตค่าศูนย์เฟ้อ (ZIG) เป็นอีกหนึ่งการแจกแจงที่นิยมที่ใช้อธิบายข้อมูลที่มีค่าศูนย์มากกว่าปกติ ในงานวิจัยนี้ ได้เสนอช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์สำหรับพารามิเตอร์ของการแจกแจงเรขาคณิตในการแจกแจงเรขาคณิตค่าศูนย์เฟ้อ โดยศึกษาในเชิงทฤษฎีและเชิงจำลอง เงื่อนไขสำหรับการหาขอบเขตล่างและบนของช่วงความเชื่อมั่นได้ถูกนำเสนอรวมถึงอสมการที่ใช้หาช่วงความเชื่อมั่น ผลจากการจำลองพบว่า ช่วงแบบโพรไฟล์ที่นำเสนอนี้ให้ความน่าจะเป็นคุ้มรวม (CP) ใกล้เคียงกับสัมประสิทธิ์ความเชื่อมั่นที่กำหนดในหลายกรณีที่ศึกษา และความยาวของช่วงโดยเฉลี่ยลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น ในบางกรณีที่มีตัวอย่างขนาดเล็ก ค่าความน่าจะเป็นคุ้มรวมยังคงใกล้เคียงกับสัมประสิทธิ์ความเชื่อมั่นที่ต้องการ

คำสำคัญ: การประมาณแบบช่วง ภาวะน่าจะเป็นโพรไฟล์ การจำลองมอนติคาร์โล ความน่าจะเป็นคุ้มรวม การแจกแจงเรขาคณิต

การอ้างอิงบทความ: พัทธ์ชนก ศรีสุรเดชชัย และ กิตติมา แดงสุภา, "ช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นโพรไฟล์สำหรับพารามิเตอร์ของการแจกแจงเรขาคณิตในการแจกแจงเรขาคณิตค่าศูนย์เฟ้อ," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 31, ฉบับที่ 3, หน้า 527–538, ก.ค.–ก.ย. 2564.

Research Article

Profile-likelihood-based Confidence Intervals for the Geometric Parameter of the Zero-inflated Geometric Distribution

Patchanok Srisuradetchai* and Kittima Dangsupa
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand

*Corresponding Author, Tel 0 2564 4444 Ext. 2101 Press 106, Email: patchanok@mathstat.sci.tu.ac.th DOI: 10.14416/j.kmutnb.2021.05.015
Received 17 September 2020; Revised 18 October 2020; Accepted 6 November 2020; Published online: 24 May 2021

Abstract

For applying statistical tools to discrete data, the frequency of zero values is sometimes greater than that of the distribution used for studies. Zero-inflated Geometric distribution (ZIG) is one of the most commonly used distributions to explain such excessive zero situations. In this study, the profile-likelihood -based confidence interval for the geometric parameter is proposed. Both theoretical and simulation studies are conducted. The conditions to obtain the lower and upper bounds of the interval are given as well as the inequality producing the interval. From the simulation study, the results suggest that profile confidence intervals yield the Coverage Probability (CP) near the given confidence coefficient in many cases of our study. The average length of the intervals decreases as the sample size increases. For some cases with small sample sizes, the CP is still close to the desirable confidence coefficient.

Keywords: Interval Estimations, Profile Likelihood, Monte-carlo Simulations, Coverage Probability, Geometric Distribution

Please cite this article as: P. Srisuradetchai and K. Dangsupa, "Profile-likelihood-based confidence intervals for the geometric parameter of the zero-inflated geometric distribution," *The Journal of KMUTNB*, vol. 31, no. 3, pp. 527–538, Jul.–Sep. 2021 (in Thai).

1. บทนำ

การแจกแจงเรขาคณิต (Geometric Distribution) เป็นการแจกแจงของตัวแปรสุ่มแบบไม่ต่อเนื่องซึ่งแทนจำนวนครั้งของการทดลองที่ล้มเหลวก่อนที่จะสำเร็จครั้งแรก อย่างไรก็ตาม การแจกแจงนี้ถูกนำไปประยุกต์ใช้ในหลายสถานการณ์ของข้อมูลที่เป็นจำนวนเต็มที่มีค่าไม่เป็นลบ เช่น นำไปอธิบายจำนวนคนพิการแต่กำเนิดของทารกแรกเกิด จำนวนครั้งที่สำรวจบ่อน้ำมันในพื้นที่ก่อนที่จะพบแหล่งที่มีศักยภาพเป็นบ่อผลิตน้ำมันได้ครั้งแรก จำนวนครั้งที่วิศวกรความปลอดภัยต้องตรวจสอบจนกว่าจะพบรายงานแสดงอุบัติเหตุที่เกิดจากการที่พนักงานไม่ปฏิบัติตามคำแนะนำ สำหรับฟังก์ชันมวลความน่าจะเป็น (Probability Mass Function; PMF) หรือพีเอ็มเอฟ เป็นดังสมการที่ (1)

$$f(x) = p(1-p)^x \qquad , x = 0,1,2,\ldots \tag{1}$$

โดยมีค่าคาดหมายและความแปรปรวนเท่ากับ $(1-p)/p$ และ $(1-p)/p^2$ ตามลำดับ และการแจกแจงเรขาคณิตเป็นการแจกแจงที่ทราบกันดีว่ามีคุณสมบัติ “Memory-less Property” ซึ่งถือว่าเป็น “Markovian Property” กล่าวคือ คุณสมบัติของตัวแปรสุ่มในอนาคตไม่ได้ขึ้นอยู่กับตัวแปรสุ่มในอดีต แต่ขึ้นกับในปัจจุบันเท่านั้น กล่าวโดยเจาะจง คือ $P(X \geq s+t \mid x \geq t) = P(X \geq s)$ [1]

ในบางสถานการณ์ ข้อมูลที่ศึกษามีค่าศูนย์มากกว่าปกติที่จะเป็นการแจกแจงเรขาคณิตแบบธรรมดา จะเรียกปัญหาที่มีลักษณะดังกล่าวนี้ว่า ปัญหาภาวะวิวิธพันธุ์ (Heterogeneity) ซึ่งโดยนัยทั่วไปจะหมายถึง สภาพที่หน่วยประชากรไม่ได้เป็นไปภายใต้ตัวแบบเดียวกัน ซึ่งอาจเป็นความต่างในรูปแบบของการแจกแจง ความแปรปรวน หรือพารามิเตอร์ต่างๆ ดังนั้น เมื่อข้อมูลมีค่าศูนย์จำนวนมากจึงมีนักสถิติเสนอการแจกแจงที่เหมาะกว่าการแจกแจงเรขาคณิต ซึ่งเรียกว่า การแจกแจงเรขาคณิตค่าศูนย์เฟ้อ (Zero-Inflated Geometric Distribution; ZIG) [2]

การแจกแจงเรขาคณิตค่าศูนย์เฟ้อนี้เป็นการแจกแจงประกอบไปด้วยพารามิเตอร์ 2 ตัว ได้แก่ พารามิเตอร์ p คือ ความน่าจะเป็นที่การทดลองแต่ละครั้งจะสำเร็จเป็นพารามิเตอร์ของการแจกแจงเรขาคณิต และพารามิเตอร์ π คือ ความน่าจะเป็นที่จะเกิดค่าศูนย์จากการแจกแจงแบร์นูลลี โดยมีพีเอ็มเอฟดังสมการที่ (2)

$$f(x;p,\pi) = \begin{cases} \pi + (1-\pi)p & , x = 0 \\ (1-\pi)p(1-p)^x & , x = 1,2,\ldots \end{cases} \tag{2}$$
$$= \left[\pi + (1-\pi)p\right]^{I_{(0)}(x)} \left[(1-\pi)p(1-p)^x\right]^{1-I_{(0)}(x)}$$

โดยที่ $0 < p < 1$, $-p/(1-p) < \pi < 1$ และ $I_{(0)}(x) = 1$ เมื่อ $x = 0$ และ $I_{(0)}(x) = 0$ เมื่อ $x \neq 0$ สามารถเขียนแทนด้วยสัญลักษณ์ $X \sim ZIG(p, \pi)$ ซึ่งมีค่าคาดหมายและค่าความแปรปรวนเท่ากับ $(1-\pi)(1-p)/p$ และ $(1-\pi)[1+\pi(1-p)](1-p)/p^2$ ตามลำดับ [3]

การแจกแจง ZIG นั้น ถูกนำเสนอครั้งแรกใน ค.ศ.1985 โดย Sharma [4] ซึ่งเพิ่มพารามิเตอร์ π หรือพารามิเตอร์การเฟ้อ (Inflation Parameter) โดยนำ ZIG นี้ไปประยุกต์ใช้กับข้อมูลการอพยพย้ายถิ่นฐานเพื่อสร้างอธิบายแนวโน้มการย้ายถิ่นออกจากชนบทสำหรับข้อมูลหมู่บ้านในประเทศอินเดีย และใช้วิธีการประมาณค่าแบบจุดด้วยวิธีโมเมนต์ (Method of Moment; MM) Iwunor [5] ได้ประมาณพารามิเตอร์โดยใช้วิธีภาวะน่าจะเป็นสูงสุด (Method of Maximum Likelihood; ML) โดยใช้ข้อมูลเดียวกันกับ Sharma [4] แต่พบว่า วิธี ML นี้ไม่มีรูปปิด (Closed Form) Aryal [6] ยังใช้การแจกแจง ZIG สำหรับข้อมูลด้านการอพยพแต่นำไปใช้กับข้อมูลของประเทศเนปาลเพื่อวางแผนนโยบายการอพยพต่างๆ อีกทั้งยังใช้ตัวประมาณแบบภาวะน่าจะเป็นสูงสุดเช่นกัน Edwin [7] ได้เสนอวิธีการประมาณค่าพารามิเตอร์แบบวิธีความถี่ของศูนย์โดยเฉลี่ย (Mean-Zero Frequency; MZF) แล้วเปรียบเทียบประสิทธิภาพกับตัวประมาณแบบ MM และ ML ผลการศึกษาพบว่า วิธี ML มีประสิทธิภาพไม่แตกต่างจากวิธี MZF โดยใช้เกณฑ์สารูปสนิทดี (Goodness of Fit Test) ในการเปรียบเทียบ

Joshi [3] ได้เสนอการแจกแจงเรขาคณิตที่มีการเฟ้อนัยทั่วไป (Generalized Inflated Geometric Distribution; GIG) ซึ่งมีกรณีพิเศษ ได้แก่ การแจกแจงเรขาคณิตค่าศูนย์เฟ้อ (ZIG) การแจกแจงเรขาคณิตค่าศูนย์และ

หนึ่งเฟ้อ (Zero-One Inflated Geometric Distribution; ZOIG) และการแจกแจงเรขาคณิตค่าศูนย์ หนึ่ง และสองเฟ้อ (Zero-One-Two Inflated Geometric Distribution; ZOTIG) และพิจารณาหาตัวประมาณพารามิเตอร์แบบจุด 2 วิธี คือ วิธี MM และ ML แล้วศึกษาโดยการจำลองเพื่อเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ ผลการศึกษาพบว่า วิธี ML จะให้ผลดีกว่าวิธี MM ในทุกการแจกแจงที่ศึกษา (ZIG, ZOIG และ ZOTIG) Mallick และ Joshi [8] ได้ประมาณพารามิเตอร์แบบจุด 3 วิธี แล้วเปรียบเทียบกับ 2 วิธี ของ Joshi [3] ซึ่งใช้วิธีฟังก์ชันก่อกำเนิดความน่าจะเป็น (Method Based on Probability Generating Function; PGF) นอกจากนี้ Kemp และ Kemp [9] พบว่า วิธี ML กับ PGF ให้ผลลัพธ์ที่ใกล้เคียงกันสำหรับการแจกแจง ZIG แต่สำหรับ ZOIG พบว่าวิธี ML มีประสิทธิภาพดีที่สุดในทุกกรณีที่ศึกษา และสำหรับการแจกแจง ZOTIG วิธี ML และ PGF มีประสิทธิภาพดีกว่า MM เมื่อขนาดตัวอย่างเพิ่มขึ้นเป็น 100 โดยผลการจำลองพบว่า วิธี ML มีประสิทธิภาพโดยรวมดีกว่าวิธี PGF และ MM

ทางด้านการวิเคราะห์การถดถอยที่เหมาะสมสำหรับข้อมูลเชิงนับ Hussein และ Hamodi [10] ศึกษาการเปรียบเทียบตัวแบบการถดถอย 3 ตัวแบบ คือ การถดถอยแบบเรขาคณิต การถดถอยแบบ Hurdle-Geometric และการถดถอยแบบ ZIG เพื่อศึกษาปัจจัยที่มีอิทธิพลต่อจำนวนผู้ติดเชื้อปอดบวมของเด็กที่มีอายุต่ำกว่า 5 ปี ในประเทศอิรัก โดยพิจารณาจากเกณฑ์ล็อกภาวะน่าจะเป็น (Log-likelihood) และเกณฑ์สารสนเทศของอะกะอิเกะ (AIC) พบว่า ตัวแบบที่เหมาะสมสำหรับข้อมูลชุดนี้ คือ ตัวแบบการถดถอยของการแจกแจง ZIG Adarabioyo และ Ipinyomi [11] ได้เปรียบเทียบตัวแบบการถดถอย 3 แบบ คือ การถดถอยจากการแจกแจงปัวซงค่าศูนย์เฟ้อ (Zero-Inflated Poisson Distribution; ZIP) การแจกแจงทวินามลบค่าศูนย์เฟ้อ (Zero-Inflated Negative Binomial Distribution; ZINB) และการแจกแจง ZIG โดยจำลองมอนติคาร์โล 1,000 รอบ กำหนดขนาดตัวอย่างที่แตกต่างกัน คือ 15, 25, 50, 100, 150, 300 และ 1000 ใช้เกณฑ์การพิจารณาจากเกณฑ์สารสนเทศของอะกะอิเกะ (AIC) และค่าคลาดเคลื่อนมาตรฐาน (Standard Error) พบว่า การแจกแจง ZIP ให้ผลดีกว่า ZINB และ ZIG โดยมี Yee [12] พัฒนาไลบรารีในโปรแกรม R ที่มีชื่อว่า VGAM (Generalized Linear and Additive Models) ที่มีฟังก์ชันที่สามารถวิเคราะห์การถดถอยแบบ ZIG ได้

Patil และ Shirke [13] ศึกษาการแจกแจงอนุกรมกำลังค่าศูนย์เฟ้อ (Zero-Inflated Power Series Distribution; ZIPS) ซึ่งเป็นกรณีทั่วไปของการแจกแจง ZIG Alshkaki [14] เสนอการประมาณค่าด้วยวิธี MM และวิธี ML สำหรับการแจกแจง ZIG และ ZOIG และนำไปประยุกต์ใช้กับข้อมูลของ Sharma [4] ต่อมา Zavaleta และคณะ [15] ได้เสนอการการทดสอบที่ขึ้นอยู่กับภาวะน่าจะเป็น (Likelihood-based Test) มา 4 วิธี ได้แก่ วิธีอัตราส่วนภาวะน่าเป็น (Likelihood Ratio) วิธีวัลด์ (Wald) วิธีราว-สกอร์ (Rao Score) และวิธีเกรเดียน (Gradient) เพื่อทดสอบสมมติฐานสำหรับ ZIPS โดยมีสมมติฐานเป็น $H_0 : \pi = 0$ และ $H_1 : \pi \neq 0$ หากสามารถปฏิเสธ สมมติฐานว่างได้แสดงว่า ข้อมูลมีการแจกแจง ZIPS มีการจำลองข้อมูลโดยใช้มอนติคาร์โล จำนวน 5,000 รอบ

ทฤษฎีอนุมานเชิงสถิติแบ่งได้เป็น 2 ส่วน คือ การประมาณค่าพารามิเตอร์ (Parameter Estimation) และการทดสอบสมมติฐาน (Hypothesis Testing) ซึ่งการประมาณค่าพารามิเตอร์จะแบ่งออกเป็น 2 วิธี คือ การประมาณค่าแบบจุด (Point Estimation) และการประมาณค่าแบบช่วง (Interval Estimation) ในงานวิจัยนี้ สนใจการประมาณแบบช่วงของพารามิเตอร์ p ในสมการที่ (2) โดยจะกำหนดสัมประสิทธิ์ความเชื่อมั่น (Confidence Coefficient) ได้ กล่าวคือ สามารถระบุระดับความคลาดเคลื่อนที่ยอมรับได้ที่เกิดจากความไม่แน่นอน (Uncertainty) ที่พารามิเตอร์ที่แท้จริงจะตกอยู่ในช่วงสุ่ม โดยแนวคิดแบบดั้งเดิมของการประมาณค่าจะขึ้นอยู่กับการแจกแจงตัวอย่าง (Sampling Distribution) ของตัวสถิติ

ซึ่งมีข้อตกลงเบื้องต้นว่า สามารถทำการทดลองซ้ำๆ ภายใต้สถานการณ์เดียวกันได้หรือที่เรียกว่า Repeated Sampling Principle [16] ในขณะที่แนวคิดแบบเบส์ (Bayesian) จะถือว่าพารามิเตอร์ที่สนใจเป็นตัวแปรสุ่มที่มีการแจกแจงความน่าจะเป็นก่อน (Prior Distribution) หลังจากที่ได้ค่าสังเกตจะหาการแจกแจงภายหลัง (Posterior Distribution) นอกจากนี้อีกแนวทางหนึ่งที่แนะนำโดย Fisher (1890–1962) คือ การอนุมานทางสถิติที่ทำโดยตรงจากฟังก์ชันภาวะน่าจะเป็น สามารถแก้จุดด้อยในการทดลองที่ไม่สามารถทำซ้ำได้ [16]

ในงานวิจัยนี้จะอิงจากแนวทางหลังสุดซึ่งการประมาณค่าแบบช่วงจะกระทำโดยตรงจากฟังก์ชันภาวะน่าจะเป็น ซึ่งมีนิยามเป็นดังสมการที่ (3)

$$L(\theta;x_{1:n}) = f(x_{1:n};\theta) \quad ,\theta\in\Theta \tag{3}$$

เมื่อ Θ แทน ปริภูมิพารามิเตอร์ (Parameter Space) และ $x_{1:n}$ แทน เวกเตอร์ของข้อมูล ฟังก์ชัน $L(\theta;x_{1:n})$ ถูกพิจารณาว่าเป็นฟังก์ชันของพารามิเตอร์ θ จะแตกต่างจาก $f(x_{1:n};\theta)$ ซึ่งถือว่าเป็นฟังก์ชันของ $x_{1:n}$ และตัวประมาณของ θ ที่ทำให้สมการที่ (3) มีค่าสูงสุดถูกเรียกว่า ตัวประมาณแบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator; MLE) ซึ่งมีนิยามเป็น $\hat{\theta}_{ML} = \underset{\theta\in\Theta}{\arg\max}\, L(\theta;x_{1:n})$ ในการอนุมานเชิงสถิตินิยมใช้ภาวะน่าจะเป็นสัมพัทธ์ (Relative Likelihood) [17] ซึ่งมีนิยามดังสมการที่ (4)

$$\tilde{L}(\theta) = \frac{L(\theta;x_{1:n})}{\max L(\theta;x_{1:n})} = \frac{L(\theta;x_{1:n})}{L(\hat{\theta}_{ML};x_{1:n})} \tag{4}$$

และช่วงแบบภาวะน่าจะเป็นสำหรับพารามิเตอร์ θ เป็นดังสมการที่ (5)

$$\left\{\theta \,\middle|\, \frac{L(\theta;x_{1:n})}{\max L(\theta;x_{1:n})} \ge c\right\} = \left\{\theta \,\middle|\, \tilde{L}(\theta) \ge c\right\} \tag{5}$$

โดยที่ $\max L(\theta;x_{1:n})$ มีค่าสูงสุดก็ต่อเมื่อ θ มีค่าเท่ากับ MLE สำหรับค่า c เป็นค่าที่นักสถิติสามารถเลือกได้ แต่โดยปกติแล้วการกำหนดค่า c นิยมใช้ทฤษฎีของวิลก์ส (Wilk's Theorem) ซึ่งระบุว่า ตัวสถิติอัตราส่วนภาวะน่าจะเป็น หรือ

$$W = -2\log_e \frac{L(\theta;x_{1:n})}{\max L(\theta;x_{1:n})} \tag{6}$$

มีการแจกแจงเชิงเส้นกำกับ (Asymptotic Distribution) เป็นการแจกแจงไคกำลังสองที่มีองศาเสรีเท่ากับ 1 แต่หาก $x_{1:n}$ มีการแจกแจงปรกติแล้ว ตัวแปรสุ่ม W ในสมการที่ (6) จะมีการแจกแจงที่แท้จริง (Exact Distribution) เป็นไคกำลังสอง ดังนั้น หาก $c = \exp\left(-\frac{1}{2}\chi^2_{1,(1-\alpha)}\right)$ แล้วช่วง (5) จะเป็นช่วงความเชื่อมั่นแบบภาวะน่าจะเป็นซึ่งเขียนได้ดังสมการที่ (7)

$$\left\{\theta \,\middle|\, \frac{L(\theta;x_{1:n})}{\max L(\theta;x_{1:n})} \ge \exp\left(-\frac{1}{2}\chi^2_{1,(1-\alpha)}\right)\right\} \tag{7}$$

โดยที่ $\chi^2_{1,(1-\alpha)}$ แทน ควอนไทล์ที่ $(1-\alpha)$ ของการแจกแจงไคกำลังสองที่มีองศาเสรีเท่ากับ 1

จากการทบทวนวรรณกรรมที่เกี่ยวข้อง การประมาณค่าแบบช่วงของพารามิเตอร์ p ในสมการที่ (2) ยังไม่มีการศึกษาในเชิงทฤษฎีโดยใช้ภาวะน่าจะเป็นโพรไฟล์ ในงานวิจัยนี้จึงมีเป้าหมายที่จะหาช่วงความเชื่อมั่นของพารามิเตอร์ของการแจกแจงเรขาคณิตในการแจกแจง ZIG โดยใช้ภาวะน่าจะเป็นโพรไฟล์กำจัดพารามิเตอร์ที่ไม่สนใจหรือ π พร้อมทั้งศึกษาหาเงื่อนไขทางคณิตศาสตร์สำหรับการสร้างช่วงความเชื่อมั่นดังกล่าว นอกจากนี้ จะศึกษาประสิทธิภาพ (Performance) ของช่วงที่นำเสนอโดยการจำลองมอนติคาร์โล (Monte-Carlo Simulations) เพื่อประมาณค่าความน่าจะเป็นคุ้มรวม (Coverage Probability; CP) และความยาวของช่วงโดยเฉลี่ย (Average Length; AL)

2. วัสดุ อุปกรณ์และวิธีการวิจัย

ในขั้นตอนการศึกษา จะแบ่งออกเป็น 2 ส่วนหลัก คือ วิธีดำเนินการวิจัยในการพิสูจน์ทางคณิตศาสตร์และในการศึกษาเชิงจำลอง

ขั้นตอนการพิสูจน์ทางคณิตศาสตร์ สามารถอธิบายได้ดังนี้

1) หาฟังก์ชันภาวะน่าจะเป็นแบบโพรไฟล์ของ p ให้

แทนด้วย $L(p,\tilde{\pi})$ โดยที่ $\tilde{\pi}$ คือ ค่าประมาณที่ทำให้ฟังก์ชันภาวะน่าจะเป็นร่วม $L(p,\pi)$ มีค่าสูงสุดเมื่อกำหนด p เป็นค่าคงที่ โดยที่ $\tilde{\pi}$ ไม่ใช่ MLE แบบปกติเพราะติดในเทอมของอีกพารามิเตอร์ (π) หนึ่ง

2) หาตัวประมาณของ p ที่ทำให้ $L(p,\tilde{\pi})$ มีค่าสูงสุด หรือ $\hat{p}^{P}_{ML} = \underset{\theta\in\Theta}{\arg\max}\, L(p,\tilde{\pi})$

3) จัดรูปหาสูตรของช่วงความเชื่อมั่น $(1-\alpha)\%$ ซึ่งมีต้นแบบดังสมการที่ (8)

$$\left\{ p \,\middle|\, \tilde{L}_p(p) \;\geq\; \exp\left(-\frac{1}{2}\chi^2_{1,(1-\alpha)}\right) \right\} \tag{8}$$

โดยที่ $\tilde{L}_p(p) = L(p,\tilde{\pi})/L(\hat{p}^{P}_{ML},\tilde{\pi})$ เรียกว่า ฟังก์ชันภาวะน่าจะเป็นโพรไฟล์สัมพัทธ์

4) เนื่องจากจะต้องมีค่า p ที่ทำให้ $\tilde{L}_p(p)$ น้อยกว่า $\exp\left(-\frac{1}{2}\chi^2_{1,(1-\alpha)}\right)$ จึงตรวจสอบว่า $\lim_{p\to 0^+}\tilde{L}_p(p)$ และ $\lim_{p\to 1^-}\tilde{L}_p(p)$ เท่ากับศูนย์หรือไม่และหาเงื่อนไข (หากมี) ที่ทำให้ลิมิตเข้าสู่ศูนย์

สำหรับวิธีศึกษาประสิทธิภาพของช่วงความเชื่อมั่นโดยการจำลองมีขั้นตอน ดังนี้

1) จำลองประชากรขนาด 1 ล้าน (ถือว่าขนาดอนันต์) ที่มีพารามิเตอร์ (p,π) แตกต่างกัน โดยที่ $\pi = -4, -2, -1, 0, 0.2, 0.4, 0.6, 0.8$ และ $p = 0.2, 0.4, 0.6, 0.8$ โดยค่า p และ π ต้องสอดคล้องกับ $-p/(1-p) < \pi < 1$ หากค่าของทั้งสองพารามิเตอร์เข้าใกล้ 1 ข้อมูลจะเป็นศูนย์ทั้งหมด ทำให้ไม่สามารถวิเคราะห์ได้ จะเห็นได้จากงานหลายชิ้น [3], [8] ในแต่ละสถานการณ์จะหาสัดส่วนของค่าสังเกตศูนย์จาก $E(N_0)/n = \pi+(1-\pi)p$ ดังสรุปในตารางที่ 1 จะสังเกตว่าทุกกรณีที่ $\pi < 0$ ประชากรจะมีศูนย์น้อยกว่าปกติหรือที่เรียกว่า Zero-deflated

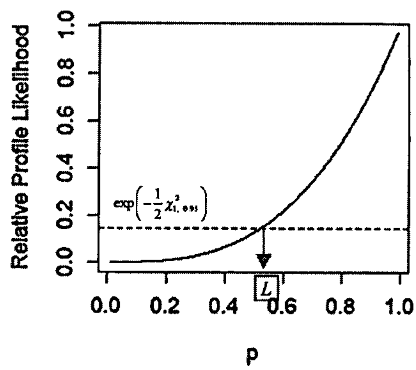
2) สุ่มตัวอย่างขนาด n จากประชากรที่จำลองขึ้น โดยกำหนด n = 10, 30, 50, 100 และ 500 ซึ่งแทนตัวอย่างขนาดเล็กไปใหญ่ตามลำดับ และมีสัมประสิทธิ์ความเชื่อมั่นเท่ากับ 0.95

3) หาขอบเขตล่าง (L) และบน (U) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับพารามิเตอร์ p จาก

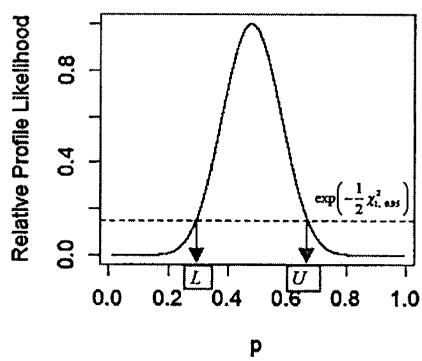
ตารางที่ 1 ค่าคาดหวังของสัดส่วนของค่าสังเกตศูนย์ใน ZIG (p, π)

π	p = 0.2	0.4	0.6	0.8
−4.0				0.00
−3.0				0.20
−2.0				0.40
−1.0			0.20	0.60
0.0	0.20	0.40	0.60	0.80
0.2	0.36	0.52	0.68	0.84
0.4	0.52	0.64	0.76	0.88
0.6	0.68	0.76	0.84	0.92
0.8	0.84	0.88	0.92	0.96

หมายเหตุ ช่องที่เว้นไว้ (p, π) ไม่อยู่ในปริภูมิพารามิเตอร์



รูปที่ 1 ฟังก์ชันภาวะน่าจะเป็นโพรไฟล์สัมพัทธ์ของตัวอย่างขนาด 100 ที่มี $\sum x_i = n_1 = 3$ จากประชากร *ZIG* $(p = 0.9, \pi = 0.8)$



รูปที่ 2 ฟังก์ชันภาวะน่าจะเป็นโพรไฟล์สัมพัทธ์ของตัวอย่างขนาด 50 ที่มี $\sum x_i > n_1$ จากประชากรจาก *ZIG* $(p = 0.4, \pi = 0.6)$

$L = \tilde{L}_p^{-1}(c)$ และ $U = \tilde{L}_p^{-1}(c)$, $U > L$ โดยที่ค่าคงที่ c เท่ากับ $\exp\left(-\chi^2_{1,(1-\alpha)}/2\right)$ ดังแสดงในรูปที่ 1 และ 2

4) ในแต่ละกรณีของข้อ 1) จะทำซ้ำข้อ 2) และ 3) จำนวน 10,000 รอบ แล้วคำนวณค่า CP จากสมการที่ (9)

$$CP = \sum_{i=1}^{10,000} I_{(L_i,U_i)}(p) \Big/ 10,000 \qquad (9)$$

โดยที่ $I_{(L_i,U_i)}(p) = 1$ เมื่อ $p \in (L_i, U_i)$ และหาก $p \notin (L_i, U_i)$ แล้ว $I_{(L_i,U_i)}(p) = 0$, $i = 1, 2, ..., 10,000$ สำหรับ AL จะคำนวณจาก

$$AL = \frac{\sum_{i=1}^{10,000} l_i}{10,000} = \frac{\sum_{i=1}^{10,000} (U_i - L_i)}{10,000} \qquad (10)$$

3. ผลการทดลอง

ผลการวิจัยจะแบ่งออกเป็น 2 ส่วน คือ เชิงทฤษฎีสถิติและเชิงการจำลอง

3.1 ผลการทดลองเชิงทฤษฎี

จะหาสูตรของ $\tilde{p}^p_{ML}$ หรือตัวประมาณที่ทำให้ฟังก์ชันภาวะน่าจะโพรไฟล์ $L\left(p,\tilde{\pi}\right)$ มีค่าสูงสุด และค่าของ $\lim_{p\to 0^+} \tilde{L}\left(p,\tilde{\pi}\right)$ และ $\lim_{p\to 1^-} \tilde{L}\left(p,\tilde{\pi}\right)$ ซึ่งเป็นการตรวจสอบว่าขอบเขตล่างและบนของช่วงสามารถหาได้หรือไม่

บทตั้ง กำหนดให้ $X_1, X_2, \ldots, X_n$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงเรขาคณิตค่าศูนย์เฟ้อที่ไม่ทราบค่าพารามิเตอร์ p และ π แล้วฟังก์ชันภาวะน่าจะเป็นโพรไฟล์ $L\left(p,\tilde{\pi}\right)$ จะมีค่าสูงสุดที่ $\tilde{p}^p_{ML} = n_1 \big/ \sum_{i=1}^{n} X_i$ โดยที่

$$\tilde{\pi} = \left(n_0 - np\right) \Big/ \left(\frac{n_0 - np}{n - np}\right)$$

โดยที่ n_0 และ n_1 แทน จำนวนค่าสังเกตในตัวอย่างที่มีค่าเป็นศูนย์และจำนวนเต็มบวก ตามลำดับ

พิสูจน์ กำหนดให้ p เป็นค่าคงที่แล้วหาตัวประมาณของ π ที่ทำให้ $L(p,\pi)$ มีค่าสูงสุดโดยการหาอนุพันธ์ของ $\log L(p,\pi)$ เทียบกับ π พิจารณา ดังนี้

$$\log L(p,\pi) = n_0 \log\left[\pi + (1-\pi)p\right] + n_1 \log p + n_1 \log(1-\pi)\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} x_i \log(1-p)$$

เมื่อหาอนุพันธ์จะได้

$$\frac{\partial}{\partial \pi}\log L(p,\pi) = \frac{n_0(1-p)}{\left[\pi + (1-\pi)p\right]} - \frac{n_1}{1-\pi}$$

กำหนด $\frac{\partial}{\partial \pi}\log L(p,\pi) = 0$ แล้วจัดรูป π ในเทอมของ p จะได้เป็นดังสมการที่ (11)

$$\tilde{\pi} = \frac{n_0 - np}{n - np} \qquad (11)$$

ดังนั้น ล็อกของฟังก์ชันภาวะน่าจะเป็นโพรไฟล์เป็น

$$\begin{aligned}
&\log L(p,\tilde{\pi}) \\
&= n_0 \log\left[\left(\frac{n_0 - np}{n - np}\right) + \left(1 - \left(\frac{n_0 - np}{n - np}\right)\right)p\right] \\
&\quad + n_1 \log\left(1 - \left(\frac{n_0 - np}{n - np}\right)\right) + n_1 \log p + \sum_{i=1}^{n} x_i \log(1-p) \qquad (12) \\
&= n_0 \log\left(\frac{n_0}{n}\right) + n_1 \log\left(\frac{n - n_0}{n - np}\right) \\
&\quad + n_1 \log p + \sum_{i=1}^{n} x_i \log(1-p)
\end{aligned}$$

แล้วหา p ที่ทำให้สมการที่ (12) มีค่าสูงสุดโดยการหาอนุพันธ์ของ $\log L\left(p,\tilde{\pi}\right)$ เทียบกับ p จะได้

$$\frac{\partial}{\partial p}\log L\left(p,\tilde{\pi}\right) = \frac{n_1}{p(1-p)} - \frac{\sum_{i=1}^{n} x_i}{1-p} \qquad (13)$$

และเมื่อกำหนดสมการที่ (13) ให้เท่ากับศูนย์ จะได้

$$\tilde{p}^p_{ML} = n_1 \Big/ \sum_{i=1}^{n} X_i$$

ทฤษฎีบท 1 กำหนดให้ $X_1, X_2, \ldots X_n$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงเรขาคณิตค่าศูนย์เฟ้อที่ไม่ทราบค่าพารามิเตอร์ π และ p แล้ว กรณีที่ $\sum x_i > n_1$

$$\lim_{p\to 0^+} \frac{L\left(p,\tilde{\pi}\right)}{\max L\left(p,\tilde{\pi}\right)} = 0 \text{ และ } \lim_{p\to 1^-} \frac{L\left(p,\tilde{\pi}\right)}{\max L\left(p,\tilde{\pi}\right)} = 0$$

และกรณีที่ $\sum x_i = n_1$ จะได้

$$\lim_{p\to 0^+}\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=0 \text{ และ } \lim_{p\to 1^-}\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=1$$

โดยที่ n_0 และ n_1 แทน จำนวนค่าสังเกตในตัวอย่างที่มีค่าเป็นศูนย์และจำนวนเต็มบวก ตามลำดับ

พิสูจน์ พิจารณา

$$\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=\frac{\exp\left(\log L(p,\tilde{\pi})\right)}{\exp\left(\log L(\tilde{p}^p_{ML},\tilde{\pi})\right)}$$

$$=\frac{\exp\left[n_0\log\left(\frac{n_0}{n}\right)+n_1\log\left(\frac{n-n_0}{n-np}\right)\right]}{\exp\left[n_0\log\left(\frac{n_0}{n}\right)+n_1\log\left(\frac{n-n_0}{n-n\tilde{p}^p_{ML}}\right)\right]}$$

$$\times\frac{\exp\left[n_1\log p+\sum_{i=1}^{n}x_i\log(1-p)\right]}{\exp\left[n_1\log\tilde{p}^p_{ML}+\sum_{i=1}^{n}x_i\log(1-\tilde{p}^p_{ML})\right]}$$

$$=\exp\left[\log\left\{\frac{\dfrac{\left(\frac{n_0}{n}\right)^{n_0}(n-n_0)^{n_1}p^{n_1}(1-p)^{\sum x_i}}{(n-np)^{n_1}}}{\dfrac{\left(\frac{n_0}{n}\right)^{n_0}(n-n_0)^{n_1}\left(\tilde{p}^p_{ML}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{\sum x_i}}{\left(n-n\tilde{p}^p_{ML}\right)^{n_1}}}\right\}\right]$$

$$=\exp\left[\log\left\{\frac{p^{n_1}(1-p)^{\sum x_i}\left(1-\tilde{p}^p_{ML}\right)^{n_1}}{(1-p)^{n_1}\left(\tilde{p}^p_{ML}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{\sum x_i}}\right\}\right]$$

$$=\left(\frac{1-\tilde{p}^p_{ML}}{\tilde{p}^p_{ML}}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{-\sum x_i}\times\frac{p^{n_1}}{(1-p)^{n_1-\sum x_i}} \quad (14)$$

กรณีที่ $\sum x_i > n_1$ จะได้ว่า

$$\lim_{p\to 0^+}\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=\left(\frac{1-\tilde{p}^p_{ML}}{\tilde{p}^p_{ML}}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{-\sum x_i}\times\lim_{p\to 0^+}p^{n_1}(1-p)^{\sum x_i-n_1}=0$$

และ $$\lim_{p\to 1^-}\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=\left(\frac{1-\tilde{p}^p_{ML}}{\tilde{p}^p_{ML}}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{-\sum x_i}\times\lim_{p\to 1^-}p^{n_1}(1-p)^{\sum x_i-n_1}=0$$

และกรณีที่ $\sum x_i = n_1$ จะได้ว่า

$$\lim_{p\to 0^+}\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=\left(\frac{1-\tilde{p}^p_{ML}}{\tilde{p}^p_{ML}}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{-\sum x_i}\times\lim_{p\to 0^+}p^{n_1}(1-p)^0$$

$$=\frac{\left(1-\tilde{p}^p_{ML}\right)^{n_1-\sum x_i}}{\left(\tilde{p}^p_{ML}\right)^{n_1}}\times\lim_{p\to 0^+}p^{n_1}=\frac{\left(1-\tilde{p}^p_{ML}\right)^{n_1-\sum x_i}}{\left(\tilde{p}^p_{ML}\right)^{n_1}}\times 0$$

ซึ่งในกรณีที่ $\sum x_i = n_1$ จะได้ว่า $\tilde{p}^p_{ML}=1$ และทำให้ $\left(1-\tilde{p}^p_{ML}\right)^{n_1-\sum x_i}=0^0$ สังเกตว่าค่าของ $\left(1-\tilde{p}^p_{ML}\right)$ และ $\left(\sum x_i - n_1\right)$ เป็นค่าที่ทราบและเป็นจำนวนเต็มศูนย์(Positive Integer) เพราะคำนวณจากตัวอย่าง มิใช่ฟังก์ชันเพื่อหาลิมิต (ที่ระบุว่าหากลิมิตเป็น 0^0 จะเป็นค่าที่มิได้กำหนดหรือ Indeterminate Forms) ในทางทฤษฎีของพหุนามนี้ หาก a เป็นจำนวนเต็มศูนย์ ค่าของ a^a จะเท่ากับ 1 [18] และในทำนองเดียวกัน

$$\lim_{p\to 1^-}\frac{L(p,\tilde{\pi})}{\max L(p,\tilde{\pi})}=\left(\frac{1-\tilde{p}^p_{ML}}{\tilde{p}^p_{ML}}\right)^{n_1}\left(1-\tilde{p}^p_{ML}\right)^{-\sum x_i}\times\lim_{p\to 1^-}p^{n_1}(1-p)^0$$

$$=\frac{\left(1-\tilde{p}^p_{ML}\right)^{n_1-\sum x_i}}{\left(\tilde{p}^p_{ML}\right)^{n_1}}\times 1=\frac{\left(1-\tilde{p}^p_{ML}\right)^{n_1-\sum x_i}}{\left(\tilde{p}^p_{ML}\right)^{n_1}}=\frac{0^0}{1}=\frac{1}{1}=1$$

จากทฤษฎีข้างต้น หาก $\sum x_i > n_1$ จะสามารถหาขอบล่างและบนของช่วงความเชื่อมั่นได้เสมอดังรูปที่ 2 แต่หาก $\sum x_i = n_1$ จะหาได้เฉพาะขอบล่างดังรูปที่ 1

ทฤษฎีบท 2 กำหนดให้ $X_1, X_2, \ldots X_n$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มมาจากประชากรที่มีการแจกแจงเรขาคณิตค่าศูนย์เฟ้อ

ที่ไม่ทราบค่าพารามิเตอร์ p และ π แล้ว ขอบล่าง (L) และขอบบน (U) ของช่วงความเชื่อมั่นเป็นรากของอสมการ

$$n_1 \log p + \left(\sum x_i - n_1\right) \log(1-p) + A \quad \geq \quad 0$$

โดยที่ $A = \left(n_1 - \sum x_i\right) \log(1 - \tilde{p}^{p}_{ML}) - n_1 \log \tilde{p}^{p}_{ML} + \chi^2_{(1-\alpha),1}/2$ และ $\tilde{p}^{p}_{ML} = n_1 / \sum x_i$ เมื่อกำหนดให้ตัวอย่างมี $\sum x_i > n_1$

พิสูจน์ จากสมการที่ (14)

$$\tilde{L}_p(p) = \frac{L\left(p, \tilde{\pi}\right)}{\max L\left(p, \tilde{\pi}\right)} = \frac{\left(1 - \tilde{p}^{p}_{ML}\right)^{n_1 - \sum x_i} p^{n_1} (1-p)^{\sum x_i - n_1}}{\left(\tilde{p}^{p}_{ML}\right)^{n_1}}$$

แล้วแทนลงในสมการที่ (8) และใส่ลอการิทึมทั้งสองข้างของอสมการแล้วจัดรูปจะได้

$$\log\left[p^{n_1} (1-p)^{\sum x_i - n_1}\right] \geq \log\left[\frac{\left(\tilde{p}^{p}_{ML}\right)^{n_1} \exp\left(-\frac{1}{2} \chi^2_{1,(1-\alpha)}\right)}{\left(1 - \tilde{p}^{p}_{ML}\right)^{n_1 - \sum x_i}}\right]$$

ซึ่งก็คือ อสมการที่กล่าวไว้

3.2 ผลการทดลองเชิงจำลอง

ค่าประมาณความน่าจะเป็นคุ้มรวม (CP) และความยาวเฉลี่ย (AL) ของช่วงความเชื่อมั่น 95% สรุปดังในตารางที่ 2 จะเห็นว่า โดยภาพรวมค่า CP มีค่าใกล้เคียงกับ 0.95 ยกเว้นในกรณีที่ p = 0.2, π = 0.8 และ n = 10 พบว่า มีค่า CP โดยประมาณ 0.8723 ซึ่งต่ำกว่า 0.95 มาก กรณีนี้เป็นกรณีที่ p มีค่าน้อยแต่ π มีค่ามากและขนาดตัวอย่างเล็กซึ่งเป็นข้อมูลที่มีศูนย์ส่วนมากแต่หากมีค่าสังเกตที่เป็นค่าบวกจะเป็นค่าที่สูง จากการสังเกตในการจำลอง การประมาณแบบช่วงมี CP ต่ำเนื่องจากช่วงที่ได้ไม่ได้กว้างมาก สำหรับในกรณีอื่นๆ ค่า CP มีแนวโน้มอยู่รอบๆ ค่า 0.95 กล่าวคือ 0.94–0.96 ในบางกรณีค่า CP มีค่าสูงกว่า 0.95 ค่อนข้างมากเช่นกรณี p เท่ากับ

ตารางที่ 2 ค่า CP และ AL ของช่วงความเชื่อมั่น

p	π	n = 10	n = 30	n = 50	n = 100	n = 500
.2	0	0.9472 (0.2629)	0.9508 (0.1458)	0.9498 (0.1123)	0.9544 (0.0788)	0.9487 (0.0351)
	0.2	0.9496 (0.2999)	0.9513 (0.1652)	0.9473 (0.1259)	0.9496 (0.0885)	0.9530 (0.0393)
	0.4	0.9523 (0.3561)	0.9480 (0.1935)	0.9489 (0.1474)	0.9474 (0.1023)	0.9487 (0.0454)
	0.6	0.9381 (0.4398)	0.9456 (0.2437)	0.9520 (0.1822)	0.9488 (0.1268)	0.9516 (0.0556)
	0.8	0.8723 (0.5605)	0.9469 (0.3633)	0.9428 (0.2724)	0.9494 (0.1842)	0.9522 (0.0793)
.4	0	0.9518 (0.4747)	0.9469 (0.2836)	0.9476 (0.2209)	0.9511 (0.1567)	0.9491 (0.0701)
	0.2	0.9605 (0.5210)	0.9456 (0.3185)	0.9475 (0.2478)	0.9490 (0.1751)	0.9494 (0.0784)
	0.4	0.9700 (0.5768)	0.9462 (0.3674)	0.9439 (0.2848)	0.9495 (0.2021)	0.9500 (0.0905)
	0.6	0.9741 (0.6454)	0.9491 (0.4435)	0.9441 (0.3498)	0.9457 (0.2474)	0.9488 (0.1108)
	0.8	0.9623 (0.7165)	0.9704 (0.5799)	0.9581 (0.4819)	0.9440 (0.3510)	0.9480 (0.1568)
.6	-1	0.9675 (0.4734)	0.9478 (0.2932)	0.9472 (0.2305)	0.9524 (0.1646)	0.9526 (0.0742)
	0	0.9837 (0.6173)	0.9530 (0.4030)	0.9487 (0.3206)	0.9475 (0.2312)	0.9537 (0.1049)
	0.2	0.9811 (0.6601)	0.9576 (0.4448)	0.9483 (0.3553)	0.9477 (0.2575)	0.9542 (0.1171)
	0.4	0.9792 (0.7049)	0.9719 (0.5017)	0.9518 (0.4049)	0.9452 (0.2962)	0.9509 (0.1354)
	0.6	0.9778 (0.7524)	0.9798 (0.5852)	0.9651 (0.4813)	0.9481 (0.3573)	0.9478 (0.1649)
	0.8	0.9784 (0.8026)	0.9812 (0.7004)	0.9789 (0.6195)	0.9680 (0.4821)	0.9441 (0.2313)
.8	-4	0.9704 (0.4052)	0.9508 (0.2428)	0.9566 (0.1920)	0.9400 (0.1379)	0.9503 (0.0625)
	-3	0.9776 (0.4552)	0.9530 (0.2696)	0.9450 (0.2133)	0.9474 (0.1541)	0.9485 (0.0698)
	-2	0.9792 (0.5231)	0.9675 (0.3087)	0.9463 (0.2438)	0.9457 (0.1769)	0.9482 (0.0807)
	-1	0.9774 (0.6177)	0.9789 (0.3762)	0.9600 (0.2956)	0.9453 (0.2145)	0.9504 (0.0986)
	0	0.9730 (0.7460)	0.9774 (0.5224)	0.9791 (0.4120)	0.9605 (0.2963)	0.9510 (0.1385)
	0.2	0.9729 (0.7741)	0.9750 (0.5745)	0.9798 (0.4588)	0.9723 (0.3289)	0.9519 (0.1542)
	0.4	0.9710 (0.8040)	0.9740 (0.6347)	0.9798 (0.5226)	0.9773 (0.3795)	0.9487 (0.1768)
	0.6	0.9690 (0.8352)	0.9770 (0.7118)	0.9743 (0.6140)	0.9783 (0.4580)	0.9447 (0.2144)
	0.8	0.9643 (0.8657)	0.9690 (0.7986)	0.9725 (0.7358)	0.9810 (0.6143)	0.9643 (0.2995)

0.6 และ 0.8, n = 10 กรณีนี้ทั้ง p และ π มีค่าค่อนข้างสูง ตัวอย่างขนาดเล็กมาก ข้อมูลส่วนใหญ่เป็นศูนย์หากมีค่าบวกจะค่าน้อยมากๆ เมื่อสังเกตผลในการจำลองพบว่า ช่วงที่ได้จะกว้างมาก ดังนั้น หากคำนวณได้ ช่วงที่ได้จะคลุมค่าพารามิเตอร์ อย่างไรก็ตาม ในทุกกรณีที่ศึกษา เมื่อขนาดตัวอย่างมีค่าเพิ่มมากขึ้น ค่า CP มีแนวโน้มลดลงเข้าสู่รอบๆ 0.95 เช่น กรณี p = 0.8 และ π = −1 ค่า CP มีค่าเท่ากับ 0.9774, 0.9789, 0.9600, 0.9453 และ 0.9504 เมื่อ n เท่ากับ 10, 30, 50, 100 และ 500 ตามลำดับ ข้อสังเกตอย่างหนึ่ง คือ กรณีที่ p = 0.8 และ π = 0.8 เป็นกรณีที่เปอร์เซ็นต์ของศูนย์มากถึง 96% จะเห็นว่า ขนาดตัวอย่างอาจจะต้องใหญ่มากจึงให้ค่า CP ใกล้เคียงกับ 0.95

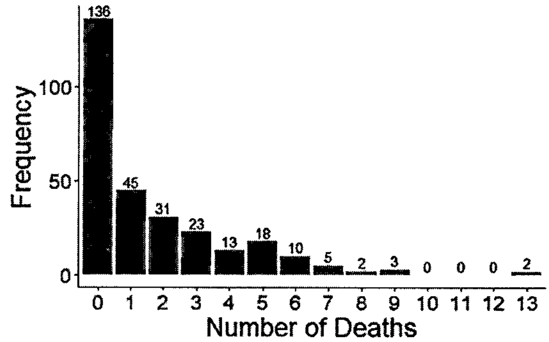
หากพิจารณาความยาวของช่วงความเชื่อมั่นแบบโพรไฟล์เมื่อเทียบกับขนาดตัวอย่างพบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า AL มีค่าลดลงในทุกกรณีที่ศึกษา ค่า AL ที่มากที่สุดมีค่าถึง 0.8657 ซึ่งเกิดในกรณีที่ p = 0.8 และ π = 0.8 (กรณีนี้มีสัดส่วนของศูนย์สูงมาก) ซึ่งสอดคล้องกับค่า AL ที่มีแนวโน้มเพิ่มขึ้นตามพารามิเตอร์ π (ของส่วนประกอบแบร์นูลลีซึ่งแสดงถึงสัดส่วนของศูนย์) เมื่อกำหนดให้ p มีค่าคงที่

4. การประยุกต์กับข้อมูลจริง

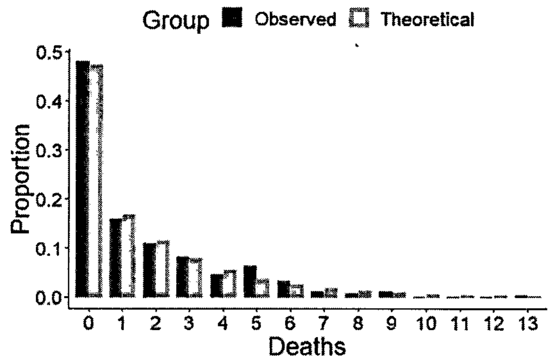
ข้อมูลผู้เสียชีวิตจาก COVID-19 ในประเทศกรีซ จากองค์การอนามัยโลก (The World Health Organization) [19] ตั้งแต่วันที่ 3 มกราคม 2563 ถึง 16 ตุลาคม 2563 รวม 288 วัน ข้อมูลแสดงดังในรูปที่ 3 จะเห็นได้ว่า ค่าสังเกตศูนย์มีจำนวนมาก เมื่อมาประมาณด้วยวิธี ML จะได้ $\hat{p}_{ML} = 0.3154$ และ $\hat{\pi}_{ML} = 0.2226$ หากนำค่าประมาณพารามิเตอร์นี้ไปประมาณค่าความน่าจะเป็น จะได้

$$P(X=x) = 0.4789^{I_{(0)}(x)}\left[0.2562(0.6704)^{x}\right]^{1-I_{(0)}(x)}$$

โดย $x = 0,1,2...$ และเมื่อนำค่าประมาณตามทฤษฎีเทียบกับสัดส่วนจริงพบว่า ค่าทั้งสองไม่ต่างกันมากดังแสดงในรูปที่ 4 กรณีนี้ ตัวอย่างมีขนาดใหญ่และมี $\sum x_i = 482, n_1 = 152$ จึงสามารถหาช่วงความเชื่อมั่นได้ตามทฤษฎีบทที่ 1 และ 2 ช่วงความเชื่อมั่น 95% แบบโพรไฟล์สำหรับจะได้เป็น (0.2749,



รูปที่ 3 จำนวนผู้เสียชีวิตจาก COVID-19 ในประเทศกรีซ

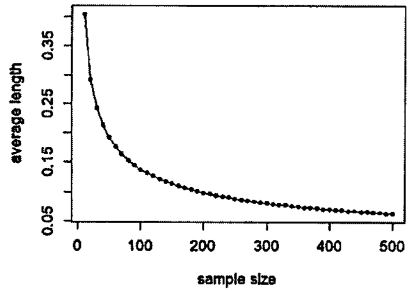


รูปที่ 4 ค่าประมาณความน่าจะเป็นและสัดส่วนผู้เสียชีวิตจริง

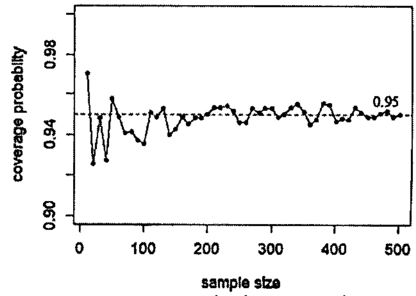
0.3577) ซึ่งมีความยาวช่วงเท่ากับ 0.0828 เมื่อเทียบกับตารางที่ 2 กรณีที่ใกล้เคียงมากที่สุด คือ p = 0.2 และ π = 0.2 เมื่อ n = 100, 500 จะมี AL เท่ากับ 0.0885 และ 0.0393 ตามลำดับ และกรณีที่ p = 0.4 และ π = 0.2 เมื่อ n = 100, 500 จะมี AL เท่ากับ 0.1751 และ 0.0784 ตามลำดับ ซึ่งค่าความยาวของช่วงที่สร้างจากข้อมูล COVID นี้ มีความเป็นไปได้ คือ ใกล้เคียงกับ 0.0784 ตามผลการจำลอง

5. อภิปรายผลและสรุป

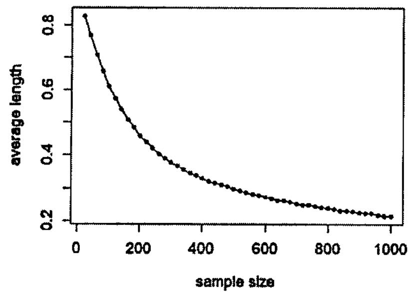
ช่วงความเชื่อมั่นแบบโพรไฟล์สำหรับพารามิเตอร์ของการแจกแจงเรขาคณิตใน ZIG ที่นำเสนอในงานวิจัยนี้ มีประสิทธิภาพดีเนื่องจาก ในหลายๆ กรณีที่ศึกษา ค่า CP เข้าใกล้ 0.95 เมื่อขนาดตัวอย่างเท่ากับเพียง 50 ในบางกรณีที่ศึกษา ตัวอย่างอาจจะต้องมีขนาดใหญ่จึงจะทำให้ ค่า CP ใกล้เคียง 0.95 เช่น กรณีที่ประชากรมี p = 0.8 และ π = −4 จะมีค่าคาดหวังสัดส่วนของค่าสังเกตศูนย์เท่ากับศูนย์



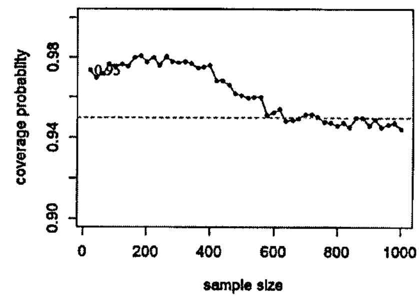
รูปที่ 5 ค่า CP ของช่วงความเชื่อมั่น 95% เมื่อประชากรเป็น $ZIG(p = 0.8, \pi = -4)$



รูปที่ 7 ค่า AL ของช่วงความเชื่อมั่น 95% เมื่อประชากรเป็น $ZIG(p = 0.8, \pi = -4)$



รูปที่ 6 ค่า CP ของช่วงความเชื่อมั่น 95% เมื่อประชากรเป็น $ZIG(p = 0.8, \pi = 0.8)$



รูปที่ 8 ค่า AL ของช่วงความเชื่อมั่น 95% เมื่อประชากรเป็น $ZIG(p = 0.8, \pi = 0.8)$

(ตารางที่ 1) ค่า CP มีค่าสูงเท่ากับ 0.97 แล้วลดลงเข้าสู่ 0.95 เมื่อขนาดตัวอย่างเพิ่ม หลังจาก $n = 200$ โดยประมาณ ค่า CP จะอยู่รอบๆ ค่า 0.95 และ AL ลดลงด้วยอัตราที่ต่ำกว่าเมื่อ $n < 200$ ดังแสดงในรูปที่ 5 และ 7 ในกรณีที่ประชากรมี $p = 0.8$ และ $\pi = 0.8$ ขนาดตัวอย่างต้องเพิ่มถึง 600 ค่า CP จึงจะอยู่รอบๆ 0.95 ในทำนองเดียวกันค่า AL หลังจาก $n = 600$ ลดลงอย่างช้าๆ ดังรูปที่ 6 และรูปที่ 8

ในทางทฤษฎีงานวิจัยนี้ได้พิสูจน์แล้วว่า ค่าขอบบนและล่างของช่วงความเชื่อมั่นหาได้เสมอหากตัวอย่างที่รวบรวมได้มี $\sum x_i > n_1$ โดยหาช่วงความเชื่อมั่นจากการแก้อสมการในทฤษฎีบท 2 ในกรณีที่ $\sum x_i = n_1$ ซึ่งมักเกิดกรณีที่ π และ p มีค่าสูงเข้าใกล้ 1 โดยจะมีค่าสังเกตที่แตกต่างกันเป็น 0 และ 1 เท่านั้น เช่น (0, 1, 0, 1, 0, 0, 0, 0, 0) จะมี $\sum x_i = n_1 = 2$ กรณีนี้ได้พิสูจน์แล้วว่า หาได้เฉพาะขอบล่างของช่วงเท่านั้น สำหรับขอบบนของช่วงอาจกำหนดให้เป็น 1 ได้เนื่องจากเป็นค่าสูงสุดของพารามิเตอร์ p

ในทางปฏิบัติอาจใช้โปรแกรม R เพื่อช่วยหาช่วงความเชื่อมั่น ในที่นี้ได้เขียนโปรแกรมแสดงในภาคผนวกซึ่งจะเป็นประโยชน์ในการวิเคราะห์ข้อมูลจริงอื่นๆ ต่อไป

6. กิตติกรรมประกาศ

ขอขอบคุณภาควิชาคณิตศาสตร์และสถิติ มหาวิทยาลัยธรรมศาสตร์ที่สนับสนุนด้านอุปกรณ์คอมพิวเตอร์

ภาคผนวก โปรแกรม R

```
# data is a vector of observed data
# 1 - alpha is level of confidence
# Default value of alpha is 0.05

profile.conf <- function(data, alpha = 0.05){
  n0 <- sum(data == 0)
  n <- length(data)
  n1 <- n - n0
  p.hat.pr <- n1/sum(data)
  A <- (n1-sum(dat))*log(1-p.hat.pr)
       -n1*log(p.hat.pr)+qchisq(1-alpha,1)/2
  func <- function(p) n1*log(p) + (sum(dat)
        -n1)*log(1-p) + A
  LB <- uniroot(func, c(0, p.hat.pr),
             tol = 1e-10)$root
  UB <- uniroot(func, c(p.hat.pr, 1),
             tol = 1e-10)$root
  return(c(LB,UB))
}
```

เอกสารอ้างอิง

[1] J. M. Horgan, *Probability with R: An Introduction with Computer Science Applications.* Hoboken, NJ: Wiley, 2020.

[2] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*. New York, NY: Cambridge University Press, 2013.

[3] R. D. Joshi, "A generalized inflated geometric distribution," M.S. thesis, College of Science, Marshall University, 2015.

[4] H. L. Sharma, "A probability distribution for rural out migration at micro level," *Rural Demography,* vol. 12, no. 1&2, pp. 63–69, 1985.

[5] C. C. O. Iwunor, "Estimating of parameters of the inflated geometric distribution for rural out-migration," *Genus,* vol. 51, pp. 3–4, 1995.

[6] T. R. Aryal, "Inflated geometric distribution to study the distribution of rural out-migrants," *Journal of the Institute of Engineering,* vol. 8, no. 1, pp. 266–268, 2011.

[7] T. K. Edwin, "Power series distributions and zero-inflated models," Ph.D. thesis, University of Nairobi, 2014.

[8] A. Mallick and R. Joshi, "Parameter estimation and application of generalized inflated geometric distribution," *Journal of Statistical Theory and Applications,* vol. 17, no.3, pp. 491–519, 2018.

[9] C. D. Kemp and A. W. Kemp, "Rapid estimation for discrete distributions," *The Statistician,* vol. 37, no. 3, pp. 243–255, 1988.

[10] M. J. M. Hussein and H. A. Hamodi, "Comparison count regression models for the number of infected of pneumonia," *Global Journal of Pure and Applied Mathematics,* vol. 13, no. 9, pp. 5359–5366, 2017.

[11] M. I. Adarabioyo and R. A. Ipinyomi, "Comparing zero-inflated poisson, zero-inflated negative binomial and zero-inflated geometric in count data with excess zero," *Asian Journal of Probability and Statistics,* vol. 4, no. 2, pp. 1–10, 2019.

[12] T. W. Yee, *Vector Generalized Linear and Additive Models: With an Implementation in R.* NY: Springer-Verlag New York, 2015.

[13] M. K. Patil and D. T. Shirke, "Testing parameter of the power series distribution of a zero inflated power series model," *Statistical Methodology,* vol. 4, pp. 393 – 406, 2007.

[14] R. S. A. Alshkaki, "Estimation of the parameters of the zero-one inflated power series distribution," *Bulletin of Mathematics and Statistics Research,* vol. 4, no. 3, 2016.

[15] K. E. C. Zavaleta, V. G. Cancho, and A. J. Lemonte, "Likelihood-based tests in zero-inflated power series models," *Journal of Statistical Computation and Simulation,* vol. 89, no. 3, pp. 443–460, 2019.

[16] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford: Clarendon Press, 2001.

[17] P. Srisuradetchai, "Profile likelihood-based confidence intervals for the mean of inverse Gaussian distribution," *The Journal of KMUTNB,* vol. 27, no. 2, pp. 339–350, 2017 (in Thai).

[18] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science.* MA: Addison-Wesley Publishing Company, 1994.

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