(2) การคำนวณหาค่าความเสี่ยงประเภทที่ 2

$$\beta(p_1) \simeq \Pr\{Z > \frac{(n/p_0 - n/p_1)}{\sqrt{nq_1/p_1^2}} + Z_\alpha \sqrt{\frac{p_1^2 q_1}{p_0^2 q_0}}\}$$

หรือ $$\beta(p_1) = \sum_{k^*+1}^{\infty} \binom{y-1}{n-1} p_1^y q_1^{y-n}$$

$$= 1 - \sum_{y=n}^{k^*} \binom{y-1}{n-1} p_1^y q_1^{y-n}$$

โดยที่ k^* คือค่าของ k' ที่คำนวณได้จากสมการ $\alpha = \sum_{y=n}^{k'} \binom{y-1}{n-1} p_0^y q_0^{y-n}$

ค. การคำนวณหาขนาดตัวอย่างอุตมะ

$$n \simeq \frac{(Z_{1-\beta} - Z_\alpha \sqrt{p_1^2 q_1/p_0^2 q_0})^2 (q_1/p_1^2)}{(1/p_0 - 1/p_1)^2}$$

สำหรับการทดสอบสมมุติฐาน $H_0 : p_0$ vs $H_1 : p \neq p_0$ จะขอเว้นไว้เป็นแบบฝึกหัด

6.2.4 การทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ p ของกลุ่มประชากรทวินาม

ก. $H_0 : p = p_0$ vs $H_1 : p = p_1 < p_0$

(1) การพัฒนาตัวทดสอบ

การทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ p ของกลุ่มประชากรทวินามมีผลเช่นเดียวกับการทดสอบเกี่ยวกับพารามิเตอร์ p ในกรณีของเบอร์นุลลิ ทั้งนี้เพราะตัวแปรสุ่มทวินามเกิดจากการทดลองเบอร์นุลลิอิสระ [1] ที่กระทำต่อเนื่องกัน n ครั้ง

ดังนั้น เมื่อทำการทดลองเบอร์นุลลิอิสระ n ครั้ง ภายใต้สมมุติฐานหลัก $H_0 : p = p_0$ และทำการทดลองเบอร์นุลลิอิสระ n ครั้ง ภายใต้สมมุติฐานรอง $H_1 : p = p_1 < p_0$. จะปรากฏ L_0 และ L_1 ดังนี้

$$L_0 \quad \prod_i^n p_0^{x_i} q_0^{1-x_i} = p_0^{\Sigma x_i} q_0^{n-\Sigma x_i}$$

และ $$L_1 = \prod_i^n p_1^{x_i} q_1^{1-x_i} = p_1^{\Sigma x_i} q_1^{n-\Sigma x_i}$$

[1] อ่านตอน 4.2.2

ดังนั้น โดยอาศัย NPL เราจะปฏิเสธสมมุติฐานหลักเมื่อ

$$\frac{L_0}{L_1} = \left(\frac{p_0}{p_1}\right)^{\Sigma x_i}\left(\frac{q_0}{q_1}\right)^{n-\Sigma x_i} \leqslant k$$

$$=> (\ln p_0 - \ln p_1)\sum_i^n x_2 - (\ln q_0 - \ln q_1) \quad x_i + n\ln q_0 - n\ln q_1 \leqslant \ln k$$

$$\{(\ln p_0 - \ln p_1) - (\ln q_0 - \ln q_1)\}\sum_i^n x_i \leqslant \ln k - n\ln q_0 + n\ln q_1$$

เนื่องจาก $p_1 < p_0$ ดังนั้น $q_1 > q_0$ ซึ่งมีผลให้ $\{(\ln p_0 - \ln p_1) - (\ln q_0 - \ln q_1)\}$ มีค่าเป็นปริมาณบวก

ดังนั้นจึงปฏิเสธสมมุติฐานหลักเมื่อ

$$\sum_i^n x_i \leqslant \frac{\ln k - n\ln q_0 + n\ln q_1}{(\ln p_0 - \ln p_1) - (\ln q_0 - \ln q_1)}$$

$$= k'$$

จากสมการ $\alpha = \Pr\{\text{Reject } H_0\}$

$=>$ $\alpha = \Pr\{\sum_i^n X_i \leqslant k' \mid p = p_0\}$

เนื่องจาก $X \sim$ Bernoulli (p) ดังนั้น ตัวแปรสุ่ม $y = \sum_i^n X_i \sim$ Binomial (n, p) และเมื่อ H_0 เป็นจริงหรือ $p = p_0$ ตัวแปรสุ่ม Y จึงมีการแจกแจงแบบ Binomial (n, p_0)

$$=> \Pr\{\sum_i^n X_i \leqslant k'\} = \alpha = \sum_{y=0}^{k'}\binom{n}{y}p_0^y q_0^{n-y}$$

หรือปฏิเสธสมมุติฐานหลักเมื่อ $y = \sum_i^n x_i \leqslant k'$ เมื่อ k' สามารถคำนวณหาได้จากสมการ

$$\alpha = \sum_{y=0}^{k'}\binom{n}{y}p_0^y q_0^{n-y}$$

อย่างไรก็ตาม เราสามารถประมาณค่าของ k' โดยอาศัย CLT ได้ดังนี้

จาก $\alpha = \Pr\{\sum_i^n X_i \leqslant k' \mid p = p_0\}$

$$=> \alpha \simeq \Pr\left\{ \frac{Y-E(Y)}{\sqrt{V(Y)}} \leqslant \frac{k'-np_0}{\sqrt{np_0q_0}} \right\} = \Pr\left\{ Z \leqslant \frac{k'-np_0}{\sqrt{np_0q_0}} \right\}$$

ดังนั้น $\Pr\left\{ Z \leqslant \frac{k'-np_0}{\sqrt{np_0q_0}} \right\} \simeq \alpha = \Pr\{Z \leqslant Z_\alpha\}$

$$=> k' \cong np_0 + Z_\alpha \sqrt{n\pi_0 q_0}$$

นั่นคือ ปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α เมื่อ

$$y = \sum_i^n x_i \leqslant np_0 + Z_\alpha \sqrt{np_0 q_0}$$

หรือปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α เมื่อ $Z_c = \frac{y-np_0}{\sqrt{np_0 q_0}} \leqslant Z_\alpha$

(2) การคำนวณหาค่าความเสี่ยงประเภทที่ 2

$$\beta(p_1) = \Pr\{\sum_i^n X_i > k^* \mid p = p_1 < p_0\}$$

ทั้งนี้ k^* คือค่าของ k' ที่คำนวณได้จากสมการ $\alpha = \sum_{y=0}^{k'} \binom{n}{y} p_0^y q_0^{n-y}$

ดังนั้น $\beta(p_1) = \sum_{y=k^*+1}^{n} \binom{n}{y} p_1^y q_1^{n-y} = 1 - \sum_{y=0}^{k^*} \binom{n}{y} p_1^y q_1^{n-y}$

หรือประมาณค่าได้โดยอาศัย CLT ดังนี้

$$\beta(p_1) \simeq \Pr\{\sum_i^n X_i > np_0 + z_\alpha \sqrt{np_0 q_0} \mid p = p_1\}$$

$$\simeq \Pr\left\{Z > \frac{np_0 - np_1}{\sqrt{np_1 q_1}} + Z_\alpha \sqrt{\frac{p_p q_0}{p_1 q_1}}\right\}$$

(3) การคำนวณหาขนาดตัวอย่างอุตมะ

จาก

$$\beta(p_1) \simeq \Pr\left\{Z > \frac{\sqrt{n}(p_0 - p_1)}{\sqrt{p_1 q_1}} + Z_\alpha \sqrt{\frac{p_0 q_0}{p_1 q_1}}\right\}$$

$$=> \sqrt{n} \simeq \frac{(Z_{1-\beta} - Z_\alpha \sqrt{p_0 q_p / p_1 q_1}) \sqrt{p_1 q_1})}{(p_0 - p_1)}$$

ดังนั้น $$n \simeq \frac{(Z_{1-\beta} - Z_\alpha \sqrt{p_0 q_0} \sqrt{p_1 q_1})^2 (p_1 q_1)}{(p_0 - p_1)^2}$$

ข. $H_0 : p = p_0$ vs $H_1 : p = p_1 > p_0$

โดยอาศัยเทคนิคและวิธีการทำนองเดียวกับ ข้อ ก. เราสามารถพิสูจน์ได้ว่า

(1) ปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α เมื่อ

$$y = \sum_{i}^{n} x_i > k'$$

โดยที่ k' สามารถคำนวณค่าได้จากสมการ

$$\alpha = \sum_{y=k'+1}^{\infty} \binom{n}{y} p_0^y q_0^{n-y}$$

$$= 1 - \sum_{y=0}^{k'} \binom{n}{y} p_0^y q_0^{n-y}$$

หรือปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α เมื่อ

$$y = \sum_{i}^{n} x_i > np_0 + Z_{1-\alpha}\sqrt{np_0 q_0}$$

หรือ

$$z_c = \frac{y - np_0}{\sqrt{np_0 q_0}} > Z_\alpha$$

(2) $\beta(p_1) = \sum_{y=0}^{k^*} \binom{n}{y} p_1^y q_1^{n-y}$

โดยที่ k^* คือค่าของ k' ที่คำนวณได้จากสมการ

$$\alpha = 1 - \sum_{y=0}^{k'} \binom{n}{y} p_0^y q_0^{n-y}$$

หรือ $$\beta(p_1) \simeq \Pr\left\{Z \leqslant \frac{np_0 - np_1}{\sqrt{np_1 q_1}} + Z_{1-\alpha}\sqrt{\frac{p_0 q_0}{p_1 q_1}}\right\}$$

(3) การคำนวณหาขนาดตัวอย่างอุตมะ

$$n \simeq \frac{(Z_\beta - Z_{1-\alpha}\sqrt{p_0 q_0 / p_1 q_1})^2 (p_1 q_1)}{(p_0 - p_1)^2}$$

สำหรับการทดสอบ $H_0 : p = p_0$ vs $H_1 : p \neq p_0$ ขอเว้นไว้เป็นแบบฝึกหัด

ตัวอย่าง 6.15 ทอดเหรียญ 100 ครั้ง ปรากฏวาเหรียญหงายก้อย 58 ครั้ง ท่านจะสรุปได้หรือไม่ว่าเหรียญนี้เป็นเหรียญถ่วงหรือไม่ถ่วง

วิธีทำ ให้ p = ความน่าจะเป็นทิเหรียญจะหงายก้อย

$H_0 : p = .5$ vs $H_1 : p > .5$

ปฏิเสธสมมุติฐานหลักเมื่อ $y > k$ โดยที่ k เป็นค่าคงที่ที่คำนวณได้จากสมการ

$$\alpha = 1 - \sum_{y=0}^{k} \binom{n}{y} p_0^y q_0^{n-y}$$

$$=> \sum_{y=0}^{k} \binom{n}{y} p_0^y q_0^{n-y} = 1 - \alpha$$

กำหนดให้ $\alpha = 0.05, n = 100$ ดังนั้น โดยอาศัยตารางทวินาม เราย่อมหาค่า k ได้ แต่เนื่องจากตารางทวินามเสนอตารางไว้สำหรับค่า n เพียง 30 (สำหรับตารางที่มีอยู่บางตารางอาจเสนอ n ถึง 50 หรือ 100) ดังนั้นเพื่อความสะดวกจึงใช้ CLT เข้าช่วย กล่าวคือ

ปฏิเสธสมมุติฐานเมื่อ

$$y = \sum_{i}^{n} x_i > np_0 + Z_{1-\alpha} \sqrt{np_0 q_0}$$

ให้ $\alpha = 0.05, n = 100, p_0 = .5, Z_{.95} = 1.64$

$$y = \sum_{i}^{100} x_i = 58$$

$$np_0 + Z_{1-\alpha} \sqrt{np_0 q_0} = 50 + (1.64) \sqrt{25} = 58.2$$

จะเห็นว่า $y = 58$ มิได้มากกว่า $np_0 + Z_{1-\alpha} \sqrt{np_0 q_0} = 58.2$ ดังนั้นจึงไม่อาจปฏิเสธสมมุติฐานหลัก และเชื่อว่าเหรียญไม่ถ่วง

หมายเหตุ เพื่อประโยชน์ในการเปรียบเทียบจะขอทดสอบสมมุติฐานนี้โดยนัยของ Goodness of fit test ดังนี้ (ดูรายละเอียดในตอนต่อไป)

$H_0 : p_1 = p_2 = \frac{1}{2}$ vs $H_1 : p_1 \neq \frac{1}{2}, p_2 \neq \frac{1}{2}$

เมื่อ p_1 = ความน่าจะเป็นที่เหรียญจะหงายหัว

p_2 = ความน่าจะเป็นที่เหรียญจะหงายก้อย

ในที่นี้ $f_X(x) = \frac{1}{2}, x = 0, 1$ (Discrete Uniform Distribution)

X_1 = จำนวนครั้งที่เหรียญหงายหัว = 42

X_2 = จำนวนครั้งที่เหรียญหงายก้อย = 58

E_1 = ค่าคาดหมายว่าเหรียญจะหงายหัว $= 100\, p_1 = 100\,(\frac{1}{2}) = 50$

E_2 = ค่าคาดหมายว่าเหรียญจะหงายก้อย $= 100\, p_2 = 10\,(\frac{1}{2}) = 50$

ปฏิเสธสมมุติฐานหลักเมื่อ $\chi_c^2 > \chi^2_{k-1,1-\alpha} = \chi^2_{1,.95} = 3.841$

$$\chi_c^2 = \sum_i^2 \frac{(X_i - E_i)^2}{E_i} = \frac{(42 - 50)^2}{50} + \frac{(58 - 50)^2}{50} = 2.56$$

จะเห็นว่า χ_c^2 ไม่ได้มากกว่า $\chi^2_{1,.95} = 3.841$ ดังนั้นจึงไม่อาจปฏิเสธสมมุติฐานหลักและเชื่อว่าเหรียญไม่ถ่วง

ตัวอย่าง 6.16 บริษัท ก. ดำริที่จะสั่งเครื่องจักรชนิดใหม่เข้ามาใช้แทนเครื่องจักรรุ่นเก่า เพราะรู้สึกว่าเครื่องรุ่นเก่าผลิตสินค้าที่ผิดมาตรฐานปะปนออกมาถึง 10% เครื่องรุ่นใหม่ที่จะสั่งซื้อนี้จะต้องมีประสิทธิภาพสูง กล่าวคือ สามารถผลิตสินค้าได้ตรงตามมาตรฐานได้มากกว่า 90%

ก. จากการทดลองนำเครื่องรุ่นใหม่มาผลิตสินค้า พบว่าเมื่อสุ่มตัวอย่างผลิตภัณฑ์ที่ผลิตโดยเครื่องจักรรุ่นใหม่มา 100 ชิ้น พบผลิตภัณฑ์ที่ผิดมาตรฐานปะปนอยู่ 12 ชิ้น ดังนี้ ถ้าท่านเป็นเจ้าหน้าที่ฝ่ายควบคุมคุณภาพ ท่านจะแนะนำบริษัทอย่างไร

ข. ถ้าจะให้การตัดสินใจของท่านผิดพลาดเพียง 1% ($\alpha = \beta = .01$) ท่านควรสุ่มตัวอย่างผลิตภัณฑ์มาตรวจสอบกี่ชิ้น ถ้า $p = .08$

วิธีทำ $H_0 : p = .10$ vs $H_1 : p < .10$

ปฏิเสธสมมุติฐานหลักเมื่อ $y \leq np_0 + Z_\alpha \sqrt{np_0 q_0}$

$n = 100, p_0 = .10$ ให้ $\alpha = .01$

ดังนั้น $np_0 + Z_\alpha \sqrt{np_0 q_0} = 10\,(2.33) \sqrt{100\,(.10)\,(.9)} = 3.01$

แต่ $y = \sum_{i}^{100} x_i = 12$ ซึ่งมิได้น้อยกว่า $np_0 + Z_\alpha\sqrt{np_0 q_0} = 3.01$

ดังนั้น จึงไม่อาจปฏิเสธสมมุติฐานหลัก และเชื่อว่าเครื่องรุ่นใหม่มิได้มีคุณภาพสูงกว่ารุ่นที่กำลังใช้อยู่

ข. ขนาดตัวอย่างอุตมะ คือ

$$n \simeq \frac{(Z_{1-\beta} - Z_\alpha\sqrt{p_0 q_0 / p_1 q_1})^2 (p_1 q_1)}{(p_0 - p_1)^2}$$

$\alpha = \beta = .01,\ p_0 = .10,\ q_0 = .90,\ p_1 = .08,\ q_1 = .92$

$$=> \quad n \simeq \frac{(2.33 + 2.33\sqrt{(.10)(.90)/(.08)(.92)})^2 (.08)(.92)}{(.10 - .08)^2}$$

$$n \simeq 4936$$

นั่นคือ ถ้าตามความเป็นจริงแล้วเครื่องรุ่นใหม่ผลิตสินค้าผิดมาตรฐานเพียง 8% และต้องการให้การตัดสินใจครั้งนี้ผิดพลาดเพียง 1% ฝ่ายควบคุมคุณภาพ ควรสุ่มตัวอย่าง ผลิตภัณฑ์ที่ทดลองผลิตโดยเครื่องรุ่นใหม่มาตรวจสอบคุณภาพประมาณ 4,936 ชิ้น

6.3 การทดสอบไคกำลังสอง (χ^2 – test)

6.3.1 บทนำ

การทดสอบไคกำลังสองโดยทั่วไปที่เราคุ้นเคยอยู่ก็คือ $\chi^2_{k-1} = \sum_{i}^{k} \frac{(O_i - E_i)^2}{E_i}$ ซึ่งนิยมใช้ทดสอบสมมุติฐานเกี่ยวกับความเป็นอิสระระหว่างคุณลักษณะทางประชากร 2 ลักษณะหรือเกินกว่า 2 ลักษณะ (Contingency Table-Test of Independence) และทดสอบความสงสัยเกี่ยวกับการแจกแจงของกลุ่มประชากรอันเป็นการตรวจสอบเพื่อหา pdf. ที่เหมาะสมของกลุ่มประชากร [1] ที่ไปเก็บตัวอย่างมา (Goodness of Fit Test) โดยปกติเป็นการนำความถี่มาใช้ แต่ธรรมชาติของความถี่จะเป็นลักษณะของการกระจายแบบตัดตอน ขณะที่การกระจายแบบ χ^2 เป็นการแจกแจงแบบต่อเนื่อง ปัญหาในชั้นต้นก็คือ χ^2-Test พัฒนามาจากแนวคิดใดและตัวทดสอบสำหรับสมมุติฐานข้างต้นพัฒนามาอย่างไร? การนำการกระจาย χ^2 มาใช้กับความถี่เหมาะสมเพียงใด? ในปัญหาที่เกี่ยวกับพัฒนาตัวทดสอบสำหรับสมมุติฐานต่าง ๆ จะได้กล่าวถึงในตอนต่อไป ในที่นี้จะได้กล่าวถึงแนวคิดเบื้องต้นอันเป็นที่มาของ χ^2-test ดังนี้

---

[1] วิธีตรวจสอบเพื่อหา pdf. ที่เหมาะสมของกลุ่มประชากรมีหลายวิธี เช่น การใช้ Probability Paper การวิเคราะห์ฮิสโตแกรม การใช้วิธี Simulation การใช้ Monte Carlo Technique การใช้ Goodness of Fit Test ฯลฯ

ถ้าตัวแปรสุ่ม $X_i \sim N(\mu_i, \sigma_i^2)$; $i = 1, 2, \ldots, n$

$$=> \quad L = \prod_{i}^{n} f_X(x_i; \mu_i, \sigma_i^2)$$

$$= = \frac{1}{\sigma_1 \sigma_2 \ldots \sigma_n (2\pi)^{n/2}} \exp\{-\frac{1}{2}\sum_{i}^{n}(\frac{x_i - \mu_i}{\sigma_i})^2\}$$

$$= \frac{1}{\prod_{i}^{n} \sigma_i (2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i}^{n} Z_i^2}$$

ซึ่งในกรณีนี้เราได้พิสูจน์มาแล้วในทฤษฎี 3.6 ว่า $Z_i^2 \sim \chi^2_{(1)}$ และ $\sum_{i}^{n}(\frac{x_i - \mu_i}{\sigma_i})^2 = \sum_{i}^{n} Z_i^2 \sim \chi^2_{(n)}$

ลองพิจารณากรณีของตัวแปรสุ่มทวินามซึ่งเป็นตัวแปรสุ่มแบบตัดตอน คือ

$$f_{X_1}(x_1; n, p_1) = \binom{n}{x_1} p^{x_1} q^{n-x_1}; \; x_1 = 1, 2, \ldots, n$$

โดยที่ p = Probability of Success และ X_1 คือ จำนวน Success จากผลการทดลองเบอร์นูลลิอิสระ n ครั้ง หรือนัยหนึ่ง X_1 คือจำนวนความถี่ (frequency) ของ p_1

ปัญหาที่พบก็คือ เราสามารถอาศัย CLT ประมาณได้ว่าตัวแปรสุ่ม $Y = \frac{X_1 - np_1}{\sqrt{np_1 q_1}}$ มีการแจกแจงแบบ $N(0, 1)$ ได้ถ้า $n \to \infty$ และ $p \to .5$ แต่ตัวแปรสุ่ม $y^2 = (\frac{X_1 - np_1}{\sqrt{np_1 q_1}})^2$ จะมีการกระจายแบบใด? สามารถอนุโลมให้มีการแจกแจงแบบ $\chi^2_{(1)}$ ได้หรือไม่?

ปัญหานี้นับเป็นประเด็นเบื้องต้นในการพัฒนา χ^2-test ซึ่งเราสามารถอาศัยความรู้เรื่อง Limiting Distribution พิสูจน์หา pdf ของ Y^2 ได้ดังนี้

ให้ $W = Y^2$

$$=> \quad F_W(w) = Pr\{W \leqslant w\}$$

$$= Pr\{Y^2 \leqslant w\}$$

$$= Pr\{-\sqrt{w} \leqslant Y \leqslant \sqrt{w}\}$$

ให้ $G_n(y)$ คือ pdf ของตัวแปรสุ่ม Y

$$=> \quad F_W(w) = Pr\{Y \leqslant \sqrt{w}\} - Pr\{Y \leqslant -\sqrt{w}\}$$

โดยอาศัย Limiting distribution $\lim\limits_{n\to\infty} G_n(y) \to N(0, 1)$

ดังนั้น $\lim\limits_{n\to\infty} F_W(w) = \lim\limits_{n\to\infty} \{Pr\, Y \leqslant \sqrt{w}\} - \lim\limits_{n\to\infty} Pr\{Y \leqslant -\sqrt{w}\}$

$$\simeq Pr\{Z \leqslant \sqrt{w}\} - Pr\{Z \leqslant -\sqrt{w}\}\ ; \text{ เมื่อ } z \sim N(0, 1)$$

$$\simeq \int_{-\infty}^{\sqrt{w}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz - \int_{-\infty}^{-\sqrt{w}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$\simeq 2 \int_{0}^{\sqrt{w}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

ดังนั้น $f_W(w) = F_{W'}(w)$

$$\simeq 2 \{\int_0^{\sqrt{w}} \frac{\partial}{\partial w} (\frac{1}{\sqrt{w\pi}} e^{-\frac{1}{2}z^2}) dz + \frac{d\sqrt{w}}{dw} (\frac{1}{\sqrt{2\pi}} 3^{-\frac{1}{2}(\sqrt{w})^2})$$

$$- \frac{d}{dw}(0) (\frac{1}{\sqrt{2\pi}} {}^{-\frac{1}{2}(0)^2})\}^1$$

$$\simeq 2\{0 + \frac{1}{2\sqrt{w}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w}\}$$

$$\simeq \frac{1}{\sqrt{w}} \quad \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w}$$

$$=> \quad f_{W'}(w) \simeq \frac{(\frac{1}{2})^2}{\Gamma(\frac{1}{2})} \cdot w^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}w}, 0 < w < \infty \ ^2$$

---

1 ดูเชิงอรรถที่ 2 หน้า 66 ในหนังสือ "ทฤษฎีสถิติ 2 เล่ม 1" โดยผู้เขียนคนเดียวกัน

2 $\because$ $W = Y^2$ และ $Y \sim N(0, 1)$ ซึ่ง $-\infty < Y < \infty$ ดังนั้น $0 < W < \infty$

นั่นคือตัวแปรสุ่ม

$$W = Y^2 \sim \chi^2_{(1)}$$

$$=> \qquad Y^2 \left(\frac{X_1 - np_1}{np_1 q_1}\right)^2 = \frac{(X_1 - np_1)^2}{np_1 q_1} \sim \chi^2_{(1)}$$

ถ้าให้ n (total frequency) $= X_1 + X_2$ โดยกำหนดให้ $p_1 p_2 = 1$ หรือ $p_2 = 1 - p_1 = q_1$ [1]

$$\text{ดังนั้น} \quad Y^2 = \frac{(X_1 - np_1)^2}{np_1 q_1} = \frac{(X_1 - np_1)^2}{np_1 (1 - p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1 - p_1)}$$

เนื่องจาก $n = X_1 + X_2 \Rightarrow X_2 = n - x_1$ และ $X_1 = n - X_2$

และ $p_1 + p_2 = 1 \Rightarrow p_1 = 1 - p_2$ ดังนั้น $np_1 = n(1 - p_2) = n - np_2$

$$\text{ดังนั้น} \quad Y^2 = \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_2 - n(1 - p_2))^2}{np_2}$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(-X_2 + np_2)^2}{np_2}$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}$$

นั่นคือ ถ้า $X_1 \sim b(n, p_1)$ โดยที่ $X_1 + X_2 = n$ และ $p_1 + p_2 = 1$ แล้วตัวแปรสุ่ม

$$Y^2 = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \sim \chi^2_{(1)}$$

ทั้งนี้ เมื่อพิจารณา $f_{X_1}(x, n, p_1)$ จะพบว่า

$$f_{X_1}(x_1, n, p_1) = \binom{n}{x_1} p_1^{x_1} q_1^{n-x_1}$$

$$= \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}$$

---

[1] X_1 = ความถี่ของ p_1, X_2 = ความถี่ของ p_2 (หรือ q_1)

$$= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} p_2^{n-x_1}$$

$\because x_1 + x_2 = n$ ดังนั้น $x_2 = n - x_1$

ดังนั้น $$f_{X_1}(x_1; n, p_1) \simeq \frac{n!}{x_1!\,x_2!} p_1^{x_1} p_2^{x_2}$$ [1]

ดังนั้น เราจึงสรุปผลได้ว่า $X_1 \sim b(n, p_1)$ และกำหนดให้ $X_1 =$ ความถี่ของ p_1 และ $X_2 =$ ความถี่ของ p_2 ($p_2 = 1 - p_1 = q_1$) ซึ่ง X_1 มี pdf ดังนี้คือ

$$f_{X_1}(x_1; n, p_1) = \frac{n!}{x_1 : x_2 :} p_1^{x_1} p_2^{x_2}$$ [2]

แล้วเราสามารถพิสูจน์ได้ว่าตัวแปรสุ่ม Y^2 เมื่อ $Y = \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}} = \frac{X_1 - np_1}{\sqrt{np_1 q_1}}$

จะมีการกระจายแบบ $\chi^2_{(1)}$ โดยประมาณเมื่อ $n \to \infty$

ทั้งนี้ $$Y^2 = \frac{(X_1 - np_1)^2}{np_1} + (\frac{(X_2 - np_2)^2}{np_2}$$

และถ้าให้ $E_1 = np_1$ และ $E_2 = np_2$

=> $$Y^2 = \frac{(X_1 - E_1)^2}{E_1} + \frac{(X_2 - E_2)^2}{E_2} = \sum_{i}^{2} \frac{(X_i - E_i)^2}{E_i} \sim \chi^2_{(1)}$$

และจากความเป็นจริงเหล่านี้ ถ้าเราขยายความสู่กรณีทั่วไปของการแจกแจงแบบพหุนาม คือเมื่อ $X_1, X_2 \ldots, X_{k-1}$ มีการแจกแจงแบบพหุนามที่มีพารามิเตอร์ $n, p_1, p_2, \ldots, p_{k-1}$ และ $X_1 + X_2 + \ldots + X_k = n$ หรือ $X_k = n\,1\,(X_1 + X_2 + \ldots + X_{k-1})$ และ $p_1 + p_2 + \ldots + p_k = 1$ หรือ $p_k = 1 - (p_1 + p_2 + \ldots + p_{k-1})$ถ้า $n \to \infty$ เราสามารถพิสูจน์ได้ว่า

$$\sum_{i}^{k} \frac{(X_i - np_i)^2}{np_i} \sim \chi^2_{(k-1)}$$

---

[1] อ่านตอน 4.2.7

[2] ขอให้สังเกตว่า $f_{X_1}(x_1; n, p_1) = \frac{n!}{x_1!\,x_2!} p_1^{x_1} p_2^{x_2}$ เป็นกรณีเฉพาะของการแจกแจงแบบพหุนาม เมื่อ $k = 2$ โดยที่ k = จำนวน พารามิเตอร์ p หรือจำนวนความถี่ X ของพารามิเตอร์ p

6.3.2 การทดสอบภาวะสารรูปสนิทดี (Goodness of Fit Test)

การทดสอบภาวะสารรูปสนิทดี คือการทดสอบดูว่า กลุ่มตัวอย่างหรือข้อมูลที่ได้รับมานั้นมีภาวะสอดคล้องกับการแจกแจงของประชากรกลุ่มที่น่าสงสัย (สารรูป) ได้ดีเพียงใด (สนิทดี) หรือนัยหนึ่งเป็นการทดสอบเพื่อเสาะหาการแจกแจงของกลุ่มประชากรอันเป็นแหล่งที่มาของกลุ่มข้อมูลนั่นเอง

ด้วยเหตุที่เราจะทราบว่าประชากรกลุ่มใดพึงมีการแจกแจงในรูปใดนั้น ดัชนีที่ใช้วัดคือความน่าจะเป็นที่ได้รับจากการกำหนดค่าหรือเซทของ Domain แล้วนำมาแจกแจงเรียกว่า Probability Distribution เช่น การแจกแจงความน่าจะเป็นที่จะมีผู้โดยสารมาลงสถานี ก. ดังนี้

จำนวนผู้โดยสารมาถึงสถานีต่อ 1 ชั่วโมง	ความน่าจะเป็น
6	.004096
7	.030720
8	.102144
9	.198400
10	.249840
11	.214200
12	.127905
13	.053550
14	.015615
15	.003100
16	.000399
17	.000030
18	.000001
	1.000000

ด้วยเหตุนี้ การทดสอบภาวะสารรูปสนิทดี จึงเป็นการทดสอบที่มุ่งเปรียบเทียบความน่าจะเป็น โดยเปรียบเทียบความน่าจะเป็นว่ามีค่าเท่ากับความน่าจะเป็นที่ได้รับจากกลุ่มประชากรที่สงสัยหรือไม่

นั่นคือ $H_0 : p_i = p_{i_0}; i = 1, 2, ..., k$

p_{i_0} = ค่าความน่าจะเป็น ณ domain ที่ $i; i = 1, 2, ..., k$ ที่คำนวณได้จากกลุ่มประชากรที่น่าสงสัย

อย่างไรก็ตาม ในการพัฒนาทางทฤษฎีนั้น เราอาจตีความหมายของสมมุติฐานได้ว่า การทดสอบที่ต้องการคือการเปรียบเทียบ Observed Frequnecy กับ Expected Frequency ที่พึงได้รับจากกลุ่มประชากรที่สงสัย เมื่อ Expected Frequency = np_{i_0}, p_{i_0} = ค่าความน่าจะเป็นที่ได้รับจากกลุ่มประชากรที่สงสัย ณ domain ที่ $i; i = 1, 2, ..., k$ และ X_i = Observed Frequency ณ domain ที่ $i; i = 1, 2, ..., k$

และขอให้สังเกตว่า $X_1, X_2, ..., X_k$ มีการแจกแจงแบบพหุนาม

การพัฒนาตัวทดสอบสามารถดำเนินการได้ดังนี้

$H_0 : p_i = p_{i_0}, i = 1, 2, ..., k$ vs $H_1\ p_i \neq p_{i_0}; i = 1, 2, ..., k$

(ก) การพัฒนาตัวทดสอบ

จากการทดลองหนึ่งที่ให้ผลลัพธ์ออกมาเป็น k อย่าง (Categories) คือ $A_1, A_2 ..., A_k$ [2] ให้ p_i คือความน่าจะเป็นที่การทดลองจะปรากฏผลลัพธ์เป็น A_i และให้ X_i คือจำนวนครั้ง (ความถี่) ของการปรากฏผลการทดลองเป็น A_i จากการทดลองอิสระ n ครั้ง ทั้งนี้ $X_1 + X_2 + ... + X_k = n$ และ $p_1 + p_2 + ... + p_k = 1$

ดังนั้นตัวแปรสุ่ม X_i ซึ่งมีการกระจายแบบทวินาม [3] หรือ $b(n, p_i)$ และ $X_1, X_2, ..., X_k$ มีการกระจายแบบพหุนาม

---

1. อ่าน Monte Carlo Technique-Isaac N. Gibra "Probability and Statistical Inference for Scientists and Engineers" p. 43.

2 เช่น ผลการทดลองทอดลูกเต๋ามี 6 อย่าง คือ หงายหน้าเอี่ยว หงายหน้า 2, ..., หงายหน้า 6 กฎของเมนเดลเสนอว่า การผสมพันธุ์พืชจะปรากฏลักษณะพืชเป็น 4 ลักษณะ ในอัตราส่วน 9 : 3 : 3 : 1

3 อาจถือว่า X_i เกิดขึ้นจากการทดลองเบอร์นุลลีที่ผลการทดลองเป็น 1 ถ้าปรากฏ A_i และมีค่าเป็น 0 ถ้าไม่ปรากฏ A_i ซึ่งในที่นี้เราจะมีตัวแปรสุ่มเป็น $X_{ij}; i = 1, 2, ..., n; j = 1, 2, ..., k$ แต่จะไม่กล่าวถึงและขอใช้วิธีพิจารณาเฉพาะตัวแปร X_i เท่านั้น จะไม่มองลึกไปถึงรากเง่าของ X_i เพราะสลับซับซ้อนกว่าแต่ให้ผลลัพธ์ตรงกัน

พิจารณา Parameter Space จะพบว่า

$\omega = \{p_{10}, p_{20}, \ldots, p_{k_0}\}$ และ $\Omega = \{p_1, p_2, \ldots, p_k; p_i \geqslant 0; i = 1, 2, \ldots, k\}$

ดังนั้น

$$L_\omega = L_\omega^{\wedge} = \prod^{k} p_{i_0}^{X_i} = p_{10}^{X_1} p_{20}^{X_2} \ldots p_{k_0}^{X_k}$$ [1]

และ

$$L_\Omega = \prod_{i}^{k} p_i^{X_i} = p_1^{X_1} p_2^{X_2} \ldots p_k^{X_k}$$

จะเห็นว่าเราไม่จำเป็นต้องประมาณค่าพารามิเตอร์ใดใน L_ω แต่ใน L_Ω จะต้องประมาณค่าพารามิเตอร์ถึง k ตัว คือ $p_1, p_2, \ldots, p_k$

แต่เมื่อพิจารณาพารามิเตอร์ $p_1, p_2, \ldots, p_k$ จะพบว่า $p_1 + p_2 + \ldots + p_k = 1$ แสดงว่าเรามีความจำเป็นต้องประมาณค่าพารามิเตอร์เพียง $k - 1$ ตัว เท่านั้น ตัวสุดท้ายที่เหลือประมาณได้จากผลต่างที่เกิดจากการนำค่าพารามิเตอร์ $k - 1$ ที่ประมาณได้ไปหักลบออกจาก 1 และเพื่อความสะดวกจะใช้ตัวที่ k คือ ถือว่า $p_k = 1 - (p_1 + p_2 + \ldots + p_{k-1}) = 1 - \sum_{i}^{k-1} p_i$

ดังนั้น เมื่อแทนที่ p_k ใน L_Ω ด้วย $1 - \sum_{i}^{k-1} p_i$ ย่อมมีผลให้มีความจำเป็นต้องประมาณค่าเฉพาะ $p_1, p_2, \ldots, p_{k-1}$ เท่านั้น

$$=> \qquad L_\Omega = p_1^{X_1} p_2^{X_2} p^{X_3} \ldots p_k^{X_{k-1}} (1 - \sum_{i}^{k-1} p_i)^{X_k}$$

$$\frac{\partial \ln l}{\partial p_i} = 0; i = 1, 2, \ldots, k - 1$$

$$=> \qquad 0 + 0 + \ldots + \frac{X_i}{p_i} + 0 + \ldots + \frac{X_k}{1 - \sum_{i}^{k-1} p_i}(-1) = 0; i = 1, 2, \ldots, k - 1$$

$$=> \qquad \frac{X_i}{p_i} - \frac{X_k}{p_k} = 0$$

ดังนั้น $$p_i = X_i \cdot \frac{p_k}{X_k}; i = 1, 2, \ldots, k - 1 \qquad \ldots\ldots\ldots(1)$$

---

[1] ในที่นี้ไม่จำเป็นต้องสนใจ $\binom{n}{x_1, x_2, \ldots, x_n}$ เพราะมิได้สนใจการสลับเทอม แต่จะให้ปรากฏเทอม $\binom{n}{x_1, x_2, \ldots, x_k}$ ไว้ด้วยก็ไม่ผิด ดู Isaaac N. Gibra, Opcit, p. 450

$\because \quad \sum_i^k p_i = 1$ ดังนั้น เมื่อรวมตลอดทุกค่าของ i ตั้งแต่ 1 ถึง k ใน สมการที่ (1)

$$=> \quad \sum_i^k \hat{p}_i = \frac{p_k}{X_k} \sum_i^k X_i$$

$$=> \quad 1 = \frac{p_k}{X_k} \cdot n \quad \because \sum_i^k X_i = n \cdot \sum_i^k \hat{p}_i = 1 = \sum_i^k p_i$$

ดังนั้น $\frac{p_k}{X_k} = \frac{1}{n}$ ....(2)

แทนค่า $\frac{p_k}{X_k}$ จากสมการ (2) ลงในสมการที่ (1)

$$=> \quad \hat{p}_1 = \frac{x_1}{n}; i = 1, 2, \ldots, k - 1$$

ขณะเดียวกัน $\hat{p}_k = 1 - (\hat{p}_1 + \hat{p}_2 + \ldots + \hat{p}_{k-1})$

$$= 1 - (\frac{X_1}{n} + \frac{X_2}{n} + \ldots \frac{X_{k-1}}{n})$$

$$= \frac{n - (X_1 + X_2 + \ldots + X_{k-1})}{n}$$

$$= X_k/n$$

ดังนั้น $L_{\hat{\Omega}} = \prod_i^k \hat{p}_i^{X_i} = (\frac{X_1}{n})^{X_1} \cdot (\frac{X_2}{n})^{X_2} \ldots (\frac{X_k}{n})^{X_k}$

โดยอาศัย MLRT เราจะปฏิเสธสมมุติฐานหลักเมื่อ $\lambda = L_{\hat{\omega}}/L_{\hat{\Omega}} \leqslant k$ นคือ ปฏิเสธสมมุติฐานหลักเมื่อ

$$\lambda = \frac{p_{10}^{X_1} p_{20}^{X_2} \ldots p_{k_0}^{X_k}}{(\frac{X_1}{n})^{X_1} (\frac{X_2}{n})^{X_2} \ldots (\frac{X_k}{n})^{X_k}} \leqslant k$$

$$= (\frac{np_{10}}{X_1})^{X_1} (\frac{np_{20}}{X_2})^{X_2} \ldots (\frac{np_{k_0}}{X_k})^{X_k}$$

จาก ทฤษฎี 7.1 เมื่อ $n \to \infty$ แล้วตัวแปรสุ่ม $-2\ln\lambda \sim \chi^2$ มี df = จำนวนพารามิเตอร์ที่กำหนดไว้ตามสมมุติฐานหลัก เมื่อ $\lambda = L_{\hat{\omega}}/L_{\hat{\Omega}}$

ดังนั้น $$-2\ln\lambda = -2\ln\{(\frac{np_{10}}{X_1})^{X_1}(\frac{np_{20}}{X_2})^{X_2}\ldots(\frac{np_{k_0}}{X_k})^{X_k}\}$$

$$= -2\sum_{i}^{k} X_i \ln(\frac{np_{i_0}}{X_i})$$

ในที่นี้ $-2\ln\lambda$ จะมีการแจกแจงโดยประมาณแบบ χ^2 มี df = k − 1 เหตุที่ df = k − 1 เพราะการแจกแจงแบบพหุนามนั้นมีข้อจำกัดว่า $\sum_{i}^{k} p_i = 1$ การกำหนดค่า p_i ในสมมุติฐานหลักจึงกำหนดเพียง k − 1 ค่าเท่านั้น ค่าที่เหลือไม่จำเป็นต้องกำหนดเพราะสามารถหาค่าได้หรือทราบค่าได้โดยอัตโนมัติ

โดยอาศัยทฤษฎี 7.1 และ MLRT เราจึงปฏิเสธสมมุติฐานหลักเมื่อ

$$-2\sum_{i}^{k} X_i \ln(\frac{np_{i_0}}{X_i}) > -2\ln k$$

กำหนดให้ $Y_i = X_i - np_{i_0}$

และเพื่อให้ง่ายขึ้นเราจะกำหนดให้ $np_{i_0} = E_i$ นั่นคือ $Y_i = X_i - E_i$ หรือ $X_i = Y_i + E_i$ [1]

$$-2\ln\lambda = -2\sum_{i}^{k}(Y_i + E_i)\ln(\frac{E_i}{Y_i + E_i}) > -2\ln k$$

$$\Rightarrow -2\ln\lambda = +2\sum_{i}^{k}(Y_i + E_i)\ln(\frac{Y_i + E_i}{E_i})$$

$$= 2\sum_{i}^{k}(Y_i + E_i)\ln(1 + \frac{Y_i}{E_i})$$

---

[1] ขอให้สังเกตว่า X_i = Observed Frequency, E_i = Expected Frequency = np_{i_0}

$$= 2 \sum_i^k (Y_i + E_i) \{\frac{Y_i}{E_i} - \frac{1}{2} (\frac{Y_i}{E_i})^2 + \frac{1}{3} (\frac{Y_i}{E_i})^3 - \ldots\}$$ [1]

$$- 2 \ln\lambda = 2 \{(\sum_i^k \frac{Y_i^2}{E_i} - \frac{1}{2} \sum_i^k \frac{Y_i^3}{E_i^2} + \frac{1}{3} \sum_i^k \frac{Y_i^4}{E_i^3} - \ldots) +$$

$$(\sum_i^k Y_i - \frac{1}{2} \sum_i^k \frac{Y_i^2}{E_i} + \frac{1}{3} \sum_i^k \frac{Y_i^3}{E_i^2} - \frac{1}{4} \sum_i^k \frac{Y_i^4}{E_i^3} + \ldots)\}$$

แต่ $$\sum_i^k Y_i = \sum_i^k (X_i - E_i) = \sum_i^k X_i - \sum_i^k E_i = n - \sum_i^k np_{i_0}$$

$$= n - n \sum_i^k p_{i_0} = n - n(1) = 0$$

ดังนั้น $$- 2 \ln\lambda = \sum_i^k \frac{Y_i^2}{E_i} - \frac{1}{3} \sum_i^k \frac{Y_i^3}{E_i^2} + \frac{1}{6} \sum_i^k \frac{Y_i^4}{E_i^3} - \ldots$$

แต่เนื่องจากตัวแปรสุ่ม $X_i \sim b(n, p_i)$ และเมื่อ H_0 เป็นจริง คือ $p_i = p_{i_0}$ ดังนั้น $X_i \sim b(n, p_{i_0})$ ซึ่งจะพบว่า $\mu_i = E(X_i) = np_{i_0} = E_i$

และ $\sigma_i^2 = V(X_i) = np_{i_0}(1 - p_{i_0}) = np_{i_0}(1 - p_{i_0}) = E_i(1 - p_{i_0})$

$$=> \qquad E_i = \sigma_i^2 / (1 - p_{i_0})$$

ดังนั้น $$\frac{Y_i}{E_i} = \frac{X_i - E_i}{E_i} = \frac{X_i - \mu_i}{\sigma_i^2 / 1 - p_{i_0}}$$

$$= \frac{X_i - \mu_i}{\sigma_i^2} \cdot (1 - p_{i_0})$$

$$= \frac{X_i - \mu_i}{\sigma_i} \cdot \frac{1 - p_{i_0}}{\sigma_i}$$

---

[1] จากอนุกรมเทย์เลอร์ของ f(u) ณ จุด a คือ

$$f(u) = f(a) + f'(a)(u - a) + f''(a)(u - a)^2 \frac{1}{2!} + \ldots + f^{(n)}(a)(u - a)^n \frac{1}{n!} + R_n$$

ดังนั้น ถ้า $f(u) = \ln(1 + u)$ เมื่อ $a = 0$ จะพบว่า

$$f(u) = \ln(1 + u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \frac{u^4}{4} + \ldots$$ เมื่อ $|u| < 1$ หรือ $-1 < u < 1$

$$= \frac{X_i - \mu_i}{\sigma_i} \cdot \frac{1 - p_{i_o}}{\sqrt{np_{i_o}}(1 - p_{i_o})}$$

$$= \frac{X_i - \mu_i}{\sigma_i} \sqrt{\frac{1 - p_{i_o}}{np_{i_o}}}$$

เมื่อ $n \to \infty$ จะพบว่า $\frac{X_i - \mu_i}{\sigma_i} \to N(0, 1)$ และ $\frac{1 - p_{i_o}}{np_{i_o}} \to 0$ นั่นคือ $n \to \infty$ แล้ว

$$\frac{Y_i}{E_i} \to N(0, 1) \text{ ที่มีค่าน้อยมาก}$$

เหตุที่ $\frac{Y_i}{E_i} \sim N(0, 1)$ ที่มีค่าน้อยมาก ดังนั้น $\frac{Y_i^2}{E_i}, \frac{Y_i^3}{E_i^2}, \frac{Y_i^4}{E_i^3}, \ldots$ ย่อมมีค่าน้อยลงตามลำดับ จนสามารถตัดทิ้งไปได้โดยมิได้มีผลกระทบกระเทือนแต่ประการใด

พิจารณา $-2 \ln\lambda = \sum_i^k \frac{Y_i^2}{E_i} - \frac{1}{3} \sum_i^k \frac{Y_i^3}{E_i} + \frac{1}{6} \sum_i^k \frac{Y_i^4}{E_i} + \ldots$

$$\Rightarrow \; -2 \ln\lambda \simeq \sum_i^k \frac{Y_i^2}{E_i}$$ [1]

นั่นคือ ปฏิเสธสมมุติฐานหลักเมื่อ $-2 \ln\lambda > -2\ln k$ หรือปฏิเสธสมมุติฐานหลัก เมื่อ

$$\sum_i^k \frac{Y_i^2}{E_i} > -2 \ln k = k'$$

จากสมการ $\alpha = \Pr\{\text{Reject } H_0 \mid H_0\}$

$$\Rightarrow \; \alpha = \Pr\{\sum_i^k \frac{Y_i^2}{E_i} > k' \mid p_i = p_{i_o}; i = 1, 2, \ldots, k\}$$

แต่ $-2 \ln\lambda \simeq \sum_i^k \frac{Y_i^2}{E_i} \sim \chi^2_{(k-1)}$

---

[1] ตัด $\sum_i^k \frac{Y_i^3}{E_i^2}, \sum_i^k \frac{Y_i^4}{E_i^3}$ และเทอมกำลังสูงอื่น ๆ ทิ้งเพราะมีค่าน้อยมาก

$$=> \Pr\{\sum_{i}^{k} \frac{Y_i^2}{E_i} > k'\} = \alpha = \Pr\{\chi^2_{(k-1)} > \chi^2_{(k-1),1-\alpha}$$

$$=> k' = \chi^2_{(k-1),1-\alpha}$$

นั่นคือ ปฏิเสธสมมุติฐานหลักเมื่อ $\sum_{i}^{k} \frac{Y_i^2}{E_i} = \sum_{i}^{k} \frac{(X_i - E_i)^2}{E_i} > \chi^2_{k-1,1-\alpha}$

โดย X_i = Observed Frequency และ $E_i = np_{i_0}$ = Expected Frequency

สำหรับการคำนวณหาความเสี่ยงประเภทที่ 2 และ Power of Test จะไม่ขอกล่าวถึง เพราะเกี่ยวข้องอยู่กับ Noncentral χ^2 Distribution นักศึกษาที่สนใจขอให้ย้อนไปดูวิธีหา β และ π ในเรื่องการทดสอบเกี่ยวกับค่าเฉลี่ยของกลุ่มประชากรปกติเมื่อไม่ทราบค่าของ σ^2

ตัวอย่าง 6.17 การบันทึกข้อมูลจากแหล่งข้อมูลซึ่งมีค่าอยู่ระหว่าง 0 – 1 โดยแบ่ง Domain ออกเป็น 4 ช่วง คือ $A_i = \{x; 0 < x < \frac{1}{4}\}$, $A_2 = \{x; \frac{1}{4} \leq x < \frac{1}{2}\}$, $A_3 = \{x; \frac{1}{2} < x \leq \frac{1}{4}$, $A_4 \{x; \frac{3}{4} < x \leq 1\}$

จากการบันทึกข้อมูล 80 ชุด ปรากฏค่าความถี่ของข้อมูลที่ตกอยู่ในช่วง A_1, A_2, A_3 และ A_4 เท่ากับ 6, 18, 20 และ 36 ชุดตามลำดับ

จงตรวจสอบดูว่าข้อมูลทั้ง 80 ชุด น่าจะเป็นข้อมูลจากกลุ่มประชากร (parent population) ใด ให้ใช้ระดับนัยสำคัญ 5%

วิธีทำ เนื่องจาก domain ของตัวแปรสุ่ม X คือ (0, 1) ดังนั้น กลุ่มประชากรที่น่าสงสัยว่าจะเป็นไปได้คือ Beta Family [1]ในที่นี้ขอตั้งข้อสงสัยเป็น 2 ประการ เพราะความถี่ค่อย ๆ ทวีขึ้น คือ

1. Triangular Distribution $f_X(x) = 2x; 0 \leq x \leq 1$ [2]
2. J-Distribution $f_X(x) = 7x^6; 0 \leq x \leq 1$

ค่าความน่าจะเป็น $p_{i_0}; i = 1, 2, 3, 4$ สำหรับ pdf. ทั้งสองพร้อมทั้งค่าความถี่คาดหมายปรากฏดังนี้

---

[1] อ่านตอน 4.2.3 โดยเฉพาะอย่างยิ่งคือ ตอน 4.2.3.1 หน้า 188

[2] ดูรูปหน้า 192

$$P_{i_0} = \int_{A_i} f_X(x)\,dx;\ i = 1, 2, 3, 4$$

และ $E_i = np_{i_0} = 80\ p_{i_0};\ i = 1, 2, 3, 4$

$P_{i_0} = \int_{A_i} 2x\,dx$	1/16	3/16	5/16	7/16
E_i	5	15	25	35
$P_{i_0} = \int_{A_i} 7x^6\,dx$	6.10351×10^{-5}	7.75146×10^{-3}	.12567138672	.8655161133
E_i	.00488	.62012	10.05371	69.32129

ก. กรณีที่สงสัยว่ากลุ่มประชากรมีการแจกแจงแบบ Triangular

$H_0 : p_1 = 1/16, p_2 = 3/16, p_3 = 5/16, p_4 = 7/16$

$H_1 : p_1 \neq 1/16, p_2 \neq 3/16, p_3 \neq 5/16, p_4 \neq 7/16$

ความถี่ที่เป็นค่าสังเกตและความถี่ที่คาดหมายว่าจะปรากฏจาก Triangular Distⁿ ณ Domain A_i; i 3 1, 2, 3, 4 ปรากฏดังตาราง

X_i	6	18	20	36
E_i	5	15	25	35

ดังนั้น $$\chi^2_c = \sum_{i}^{4} \frac{(X_i - E_i)^2}{E_i}$$

$$= \frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35}$$

$$= 1.828$$

จากตาราง $\chi^2_{3,.95} = 7.815$

จะเห็นว่า $\chi^2_c < \chi^2_{3,.95} = 7.815$

ดังนั้น จึงไม่อาจปฏิเสธสมมุติฐานหลักหรือเชื่อว่ากลุ่มประชากรที่สงสัยว่าเป็น Parent Population คือ Triangular Dist$^{\underline{n}}$ หรือเชื่อว่าค่าความน่าจะเป็น p_i สอดคล้องกับค่า p_{i_0} ที่ได้รับจาก Triangular Distribution

อนึ่ง ขอให้เป็นที่สังเกตว่า ตามตัวอย่างนี้ X_i และ E_i มีค่าใกล้เคียงกันมาก ณ จุดนี้เราจึงพอจะเดาคำตอบได้จากการเปรียบระหว่างค่าสังเกตกับค่าคาดหมายทีละคู่ ถ้าหากมีค่าใกล้เคียงกันก็พอจะเชื่อได้ว่าสมมุติฐานหลักถูกต้อง ถ้าหากมีค่าแตกต่างกันมากเราเชื่อว่าสมมุติฐานรองน่าจะถูกต้อง

ข. กรณีที่น่าสงสัยว่ากลุ่มประชากรมีการแจกแจงแบบ j

H_0: $p_1 = 6.10351 \times 10^{-5}$, p_2 $7.75146 \times ^{-3}$, pa = .12567, p_4 = .86551

H_1 : $p_1 \neq .10351 \times 10^{-5}$, $p_2 \neq 7.75146 \times 10^{-3}$, $p_3 \neq .12567$, $p_4 \neq .86551$

เปรียบเทียบความถี่ได้ดังตาราง

X_i	6	18	20	36
E_i	.00488	.62012	10.05371	69.32129

ก่อนที่จะวิเคราะห์ตัวอย่างนี้ขอให้นักศึกษาพึงทราบข้อจำกัดของ χ^2 – test ดังนี้ [1]

เนื่องจาก $\sum_i^k \frac{(X_i - E_i)^2}{E_i}$ มิใช่มีการแจกแจงแบบ χ^2 โดยตรง แต่เป็นเพียงมีการแจกแจงแบบ χ^2 โดยประมาณ ดังนั้น เพื่อให้ $\sum_i^k \frac{(X_i - E_i)^2}{E_i}$ มีการแจกแจงใกล้เคียงกับ χ^2 – Dist– ขนาดตัวอย่าง n ควรมีขนาดใหญ่มาก ๆ เพื่อให้ $E_i \geqslant 5$ ในขณะเดียวกันจำนวน k [2] จะต้องไม่น้อยกว่า 5 คือ $k \geqslant 5$ หรือถ้า $k < 5$ ค่าของ E ก็จะต้องสูงกว่า 5 มิฉะนั้น การประมาณค่าด้วย χ^2 – dist$^{\underline{n}}$ จะไม่สัมฤทธิผล

---

[1] Hoel, Opcit., p. 247 Isaac N. Gibra, Opcit., p. 453, Hogg and Craig, Opcit, p. 309.

[2] k = จำนวนตัวแปรสุ่ม X ในฟังก์ชันการแจกแจงพหุนาม หรือในทางปฏิบัติ k = จำนวนกลุ่มย่อยของ domain หรือจำนวน category

หนทางสำหรับแก้ปัญหา เมื่อ $E_i < 5$ ให้ปฏิบัติดังนี้

1) รวมค่าความถี่คาดหมายในกลุ่มที่อยู่ใกล้เคียงกันเข้าด้วยกันจนกระทั่งทำให้ $E_i \geq 5$

2) รวมค่าความถี่สังเกตที่สอดคล้องกับความถี่ใน ข้อ 1) เข้าด้วยกัน

3) df ของ $\sum_{i}^{r} \frac{(X_i - E_i)^2}{E_i}$ มีค่าเท่ากับ $r - 1$ เมื่อ r คือจำนวน category ที่เหลืออยู่หลังจากการดำเนินการใน ข้อ 1) และ 2)

อนึ่ง ค่า E_i ควรเป็นเลขจำนวนเต็มเพราะเป็นค่าความถี่ แต่ในหลายกรณีค่า E_i มักมิได้เป็นเลขจำนวนเต็ม ซึ่งในกรณีเช่นนี้ยังคงให้ใช้ค่านั้นได้ต่อไป เพราะการปรับค่าให้กลับเป็นเลขจำนวนเต็มอาจมีผลเสียต่อค่าของ χ_c^2 และโครงสร้างเดิมอันเป็นผลลัพธ์ของการพัฒนาตัวทดสอบ

ตามตัวอย่างใน ข้อ ข. เราสามารถปรับตารางเสียใหม่โดยนำความถี่ X_1 และ X_2 รวมกับ X_3 และนำ E_1 และ E_2 รวมกับ E_3 ซึ่งปรากฏผลดังตารางต่อไปนี้

X_i	44	36
E_i	10.67871	69.32129

$$\chi_c^2 = \frac{(44 - 10.67871)^2}{10.67871} + \frac{(36 - 69.32129)^2}{69.32129} \sim \chi^2_{(1)}$$

$$= 119.99086$$

ขณะที่ $\chi^2_{1,.95} = 3.481$

เราจึงสรุปได้ว่าค่า p_i มิได้สอดคล้องกับค่า p_{i_0} ที่ได้รับมาจาก J-Distribution หรืออีกนัยหนึ่งเชื่อว่ากลุ่มตัวอย่างนี้มิได้เป็นกลุ่มตัวอย่างของประชากร J-Distribution

ตัวอย่าง 6.18 ในการบันทึกข้อมูลจากแหล่งข้อมูลที่ให้ค่าอยู่ในช่วง $(-\infty, \infty)$ โดยแบ่ง domain ออกเป็น 8 ช่วง คือ $A_1 = \{x; -\infty < x < 0\}$, $A_i = \{x_i; i-2 < x \leq i-1\}; i = 1, 2, \ldots, 7$ และ $A_8 = \{x; 6 < x < \infty\}$ ปรากฏความถี่ของข้อมูลที่ปรากฏในช่วง $A_1, A_2, \ldots, A_8$ เท่ากับ 60, 96, 140, 210, 172, 160, 88 และ 74 ชุด ตามลำดับ

อยากทราบว่าข้อมูลทั้ง 1,000 ชุด นี้น่าจะสุ่มมาจากกลุ่มประชากรใดหรือ Parent Population น่าจะมีการแจกแจงแบบใด ให้ใช้ระดับนัยสำคัญ 5%

วิธีทำ จากการพิจารณา domain จะพบว่าค่าของตัวแปรสุ่มปรากฏในช่วง $(-\infty, \infty)$ อนึ่ง เมื่อพิจารณาความถี่ในช่วงต่าง ๆ พบว่ามีความถี่เท่ากับ 60, 96, 140, 210, 172, 160, 88, 74 โดยลักษณะของฮิสโตแกรมแล้วเข้าข่ายโค้งปกติ ด้วยเหตุนี้เราจึงเดาได้ว่า Parent Population น่าจะมีการแจกแจงแบบปกติ

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}; -\infty < x < \infty$$

และเพื่อความสะดวกขอถือว่า $\mu = 3, \sigma^2 = 4$ [1]

ดังนั้น

$$p_{i_0} = \int_{A_i} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\, dx; i = 1, 2, \ldots, 8$$

และความถี่คาดหมาย $E_i = np_{i_0} = 1000\, p_{i_0}; i = 1, 2, \ldots, 8$

สำหรับการคำนวณหาค่า p_{i_0} ให้ปฏิบัติดังนี้

$$p_{1_0} = \int_{-\infty}^{0} \frac{1}{\sigma\sqrt{2\pi}}\, e^{1/2(x-\mu)^2/\sigma^2}\, dx$$

$$= \int_{-\infty}^{0} \frac{1}{2\sqrt{2\pi}}\, e^{-\frac{1}{2(4)}(x-3)^2}$$

[1] ในทางปฏิบัติเราอาจประมาณค่า υ และ σ^2 จากกลุ่มตัวอย่างหรืออาศัยข้อสนเทศในอดีต

ให้ $\frac{x-3}{2} = z$ ดังนั้น $dx = 2\,dz$ และ $-\infty \leqslant z \leqslant -1.5$

$$=> p_{1_0} = \int_{-\infty}^{-1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = \int_{-\infty}^{-1.5} f_z(x)\,dz$$

$$= \Pr\{Z \leqslant -1.5) = .06681$$

ในทำนองเดียวกัน

$$p_{20} = \int_{-1.5}^{-1} f_z(z)\,dz = \Pr\{Z \leqslant -1\} - \Pr\{Z \leqslant -1.5)$$

$$= .1587 - .06681 = .09189$$

$$p_{30} = \int_{-1}^{-.5} f_Z(z)\,dz = \Pr\{Z \leqslant -.5\} - \{\Pr Z \text{ 6 } -1\} = .3085 - .1587$$

$$p_{40} = \int_{-.5}^{0} f_z(z)\,dz = \Pr\{Z \leqslant 0\} - \Pr\{Z \leqslant -.5\} = .5000 - .3085$$

$$p_{50} = \int_{0}^{.5} f_Z(z)\,dz = \Pr\{Z \leqslant .5\} - \Pr\{Z \leqslant 0) = .6915 - .5000$$

$$p_{60} = \int_{.5}^{1} f_Z(z)\,dz = \Pr\{Z \leqslant 1\} - \Pr\{Z \leqslant .5\} = .8413 - .6915$$

$$p_{70} = \int_{1}^{1.5} f_Z(z)\,dz = \Pr\{Z \leqslant 1.5\} - \Pr\{Z \leqslant 1) = .93319 - .8413$$

$$p_{80} = \int_{1.5}^{\infty} f_,(z)\,dz = \Pr\{Z < \infty\{ - \Pr\{Z \leqslant 1.5) = 1 - .93319$$

ดังนั้น ค่า p_{i_0} และ E_i; $i = 1, 2, \ldots, 8$ ปรากฏดังตารางต่อไปนี้

Pi.	.06681	.09189	.1498	.1915	.1915	.1498	.09189	.06681
E_i	66.81	91.89	149.8	191.5	191.5	149.8	91.89	66.81

ดังนั้น $H_0: p_1 = .06681, p_2 = .09189, p_3 = .1498, p_4 = .1915, p_5 = .1915$

$p_6 = .1498, p_7 = .09189, p_8 = .06681$

$H_1: p_1 \neq .06681, \ldots, p_7 \neq .09189, p_8 \neq .06681$

ตารางเปรียบเทียบระหว่างความถี่สังเกตกับความถี่คาดหมาย ปรากฏดังนี้

X_i	60	96	140	210	172	160	88	74
E_i	66.81	91.89	149.8	191.5	191.5	149.8	91.89	66.81

$$\chi_c^2 = \frac{(60 - 66.81)^2}{66.81} + \frac{(96 - 91.89)^2}{91.89} + \ldots + \frac{(74 - 66.81)^2}{66.81}$$

$$= 6.92492626262264$$

$$\chi^2_{7,.95} = 14.07$$

จะเห็นได้ว่า $\chi_c^2 < 14.07$

นั่นคือ เราไม่อาจปฏิเสธสมมุติฐานหลักและเชื่อว่ากลุ่มประชากรมีการแจกแจงแบบปกติ N (3, 4)

ตัวอย่าง 6.19 ในการศึกษากลุ่มเลือดพบว่ามนุษย์จะมีอัตราของกลุ่มเลือดทั้ง 4 กลุ่ม คือ A, B, AB และ 0 ในอัตราส่วน $q^2 : p^2 + 2pq : r^2 + 2qr : 2pr$

จากการเจาะเลือดของคนทั่วไปโดยสุ่ม 770 คน พบว่าเป็นผู้มีกลุ่มเลือด A รวม 180 คน กลุ่มเลือด B รวม 360 คน กลุ่มเลือด AB รวม 132 คน กลุ่มเลือด 0 รวม 98 คน

อยากทราบว่าค่าของความถี่ข้างต้นนี้สนับสนุนอัตราส่วน $p = .4, q = .4, r = .2$ หรือไม่?

วิธีทำ ถ้า $p = .4, q = .4, r = .2$ ดังนั้น

$$q^2 : p^2 + 2pq : r^2 + 2qr : 2pr = 0.16 : 0.48 : 0.20 : 0.16$$

นั่นคือ $H_0: p_1 = 16/100, p_2\ 48/100, p_3\ 20/100, p_4 = 16/100$

$H_1: p_1 \neq 16/100, p_2 \neq 48/100, p_3 \neq 20/100, p_4 \neq 16/100$

ตารางเปรียบเทียบระหว่างความถี่สังเกตและความถี่คาดหมายปรากฏดังนี้

Xi	180	360	132	98
$E_i = np_{i_o} = 770\ p_{i_o}$	123.2	369.6	154	123.2

$$\chi_c^2 = \frac{(180 - 123.2)^2}{123.2} + \frac{(360 - 369.6)^2}{369.6} + \frac{(132 - 154)^2}{154} + \frac{(98 - 123.3)^2}{123.2}$$

$$= 34.73376623371$$

แต่ $\chi^2_{3,.95} = 7.815$ ซึ่งจะพบว่า $\chi_c^2 > 7.815$

นั่นคือ เราปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ 5% และเชื่อว่าทฤษฎีเกี่ยวกับกลุ่มเลือดที่ว่าอัตราส่วนระหว่างคนที่มีกลุ่มเลือด A, B, AB และ 0 ที่เท่ากับ $q^2:p^2+2pq:r^2+2qr:2pr$ นั้น ค่า p, q, r ไม่ควรมีค่าเป็น .4, .4 และ .2

ตัวอย่าง 6.20 จากการบันทึกจำนวนเรือบรรทุกน้ำมันที่เข้าโรงกลั่นรวมทั้งสิ้น 1,000 วัน ปรากฏข้อมูลดังนี้

จำนวนเรือที่เข้าโรงกลั่นต่อวัน	ความถี่
0	376
1	356
2	191
3	57
4	16
5	2
6	1
7	1
	1,000

อยากทราบว่าจำนวนเรือที่เข้าโรงกลั่นในแต่ละวันเป็นตัวแปรสุ่มพัวซองหรือไม่ ทั้งนี้ให้ใช้ระดับนัยสำคัญ 5%

วิธีทำ $H_0 : p_i = p_{i_0}; i = 0, 1, 2, ..., 7$

เหตุที่ λ = อัตราถั่วเฉลี่ยจำนวนเรือที่เข้าโรงกลั่นต่อวัน

ดังนั้น $\lambda = \frac{1}{1,000}(0 \times 376 + 1 \times 356 + 2 \times 191 + ... + 7 \times 1)$

$= 1$ ลำ/วัน

ดังนั้น $p_{i_0} = \frac{e^{-1} 1^y}{y!} = \frac{1}{ey!}; y = 0, 1, 2, ..., 7$

จากตารางพัวซองพบว่า

$p_{00} = .368, p_{10} = .368, p_{20} = .184, p_{30} = .061$

$p_{40} = .015, p_{50} = .003, p_{60} = .001, p_{70} = 0$

$E_i = np_{i_0} = 1,000\, p_{i_0}; i = 0, 1, 2, ..., 7$

ดังนั้น จึงปรากฏข้อมูลเปรียบเทียบระหว่างค่าสังเกตและค่าคาดหมายดังนี้

จำนวนเรือ	0	1	2	3	4	5	6	7
X_i	376	256	191	57	16	2	1	1
E_i	368	368	184	61	15	3	1	0

จัดกลุ่มข้อมูลค่าคาดหมาย E_i ที่มีค่าต่ำกว่า 5 เข้าเป็นกลุ่มที่มีค่าอย่างน้อยเท่ากับ 5 ในที่นี้กระทำโดยนำ 3 และ 1 ไปรวมกับ 15 ข้อมูลที่ต้องการจึงปรากฏดังตาราง

จำนวนเรือ	0	1	2	3	4 - 7
X_i	376	356	191	57	20
E_i	368	368	184	61	19

$$\chi_c^2 = \frac{(376-368)^2}{368} + \frac{(356-368)^2}{368} + \frac{(191-184)^2}{184} + \frac{(57-61)^2}{61} + \frac{(20-19)^2}{19} = 1.146$$

แต่ $\because \chi^2_{4,.95} = 9.48$

จะเห็นว่า $\chi^2_c = 1.146$ มิได้มากกว่า $\chi^2_{4,.95}$ ดังนั้น จึงไม่อาจปฏิเสธสมมุติฐานหลัก และเราเชื่อว่าจำนวนเรือที่เข้าโรงกลั่นเป็นตัวแปรสุ่มพัวซอง

6.3.3 ตารางจำแนก (Contingency Table : CT)

CT คือ ตารางจำแนกหลายทาง (Multiple Classification) หมายความว่า จำแนกกลุ่มตัวอย่างออกเป็น 2 ทาง 3 ทาง 4 ทาง หรือมากกว่านั้นตามคุณลักษณะทางประชากรที่สนใจ โดยมีเจตนาเพื่อศึกษาความเกี่ยวข้องสัมพันธ์กันในระหว่างตัวแปรที่แสดงคุณลักษณะทางประชากร และโดยปกติเราจะจำแนกคุณลกษณะทางประชากรหนึ่ง ๆ ออกเป็นหลายระดับ ดังนั้น เราจึงอาจกล่าวได้ว่า CT คือตารางจำแนกที่มุ่งตรวจสอบดูว่าระดับต่าง ๆ ของคุณลักษณะหนึ่งของกลุ่มประชากรจะผันผวนไปตามระดับต่าง ๆ ของอีกคุณลักษณะหนึ่งของประชากรหรือไม่ ตัวอย่างเช่น ผู้วิจัยอยากทราบว่าความพิถีพิถันในการเลือกคู่ครองผันผวนไปตามระดับการศึกษาของผู้เลือกคู่ครองหรือไม่ ในกรณีเช่นนี้เราอาจจำแนกความพิถีพิถันออกเป็น 4 ระดับ คือ ไม่พิถีพิถันเลย พิถีพิถันน้อย พิถีพิถันมาก และพิถีพิถันที่สุด ขณะที่จำแนกระดับการศึกษาของบุคคลออกเป็น 4 ระดับ คือ ไม่มีการศึกษา มีการศึกษา ระดับประถมศึกษา มีการศึกษาระดับมัธยมศึกษา และมีการศึกษาระดับอุดมศึกษา การจำแนกลักษณะนี้เรียกว่า 4×4 CT หรือผู้วิจัยต้องการทราบว่า ความพึงพอใจในกฎหมายรัฐธรรมนูญของประชาชนผันแปรไปตามเพศและระดับรายได้ของประชากรหรือไม่ ในกรณีเช่นนี้อาจจำแนกความพึงพอใจในกฎหมายรัฐธรรมนูญออกเป็น 3 ระดับ คือ พอใจ ไม่พอใจ และเฉย ๆ ขณะที่จำแนกเพศออกเป็นเพศชายและเพศหญิง และจำแนกระดับรายได้ต่อเดือนเป็น 4 ระดับ คือ ไม่เกิน 5,000 บาท 5,000 - 10,000 บาท 10,000 - 15,000 บาท และ 15,000 บาท ขึ้นไป ลักษณะนี้เรียกว่า $3 \times 2 \times 4$ CT ดังนี้เป็นต้น จะเห็นได้ว่า CT จะเป็นตารางที่แสดงหรือแจกจำแนกกลุ่มตัวอย่างออกเป็นกลุ่มย่อย ๆ $r \times c$ กลุ่ม หรือ $r \times c \times k$ กลุ่ม เช่น ตามตัวอย่างข้างต้น ตาราง CT จำแนกกลุ่มตัวอย่างเป็น $4 \times 4 = 16$ กลุ่มย่อย หรือ $3 \times 2 \times 4 = 24$ กลุ่มย่อย เรียกกลุ่มย่อย ๆ เหล่านี้ว่า ห้อง (cell) ดังนั้น แต่ละ cell จึงแสดงความถี่หรือจำนวนหน่วยตัวอย่างที่มีคุณสมบัติสอดคล้องกับระดับย่อยที่สอดคล้องกัน ดังตัวอย่างเรื่องการศึกษาความสัมพันธ์ระหว่างทัศนคติต่อกฎหมายรัฐธรรมนูญกับเพศของประชาชน สมมุติว่าทำการสำรวจกับประชาชน 3,750 คน ปรากฏข้อมูลดังตาราง

เพศ	ทัศนคติที่มีต่อกฎหมายรัฐธรรมนูญ		
	พอใจ	ไม่พอใจ	เฉย ๆ
ชาย	1155	475	243
หญิง	1083	442	362

ข้อมูลใน cell ที่ (1, 1) คือ 1,155 แสดงจำนวนกลุ่มตัวอย่างย่อย (ความถี่) ที่เป็นเพศชายและพึงพอใจในกฎหมายรัฐธรรมนูญ หรือใน cell ที่ (2, 3) ข้อมูลคือ 362 แสดงจำนวนกลุ่มตัวอย่างย่อย (ความถี่) ที่เป็นเพศหญิง และไม่มีความรู้สึกใด ๆ ต่อกฎหมายรัฐธรรมนูญ เป็นต้น

ดังนั้น สิ่งที่สนใจศึกษาจึงมุ่งไปที่การศึกษาว่าคุณลักษณะทางประชากรมีความเกี่ยวเนื่องกันหรือไม่ สมมุติฐานหลักที่ใช้จึงเสนอไว้ในลักษณะเป็นกลาง คือคุณลักษณะทางประชากรที่สนใจนั้นไม่เกี่ยวข้องกันหรือเป็นอิสระต่อกัน ขณะที่สมมุติฐานรองหรือสมมุติฐานเพื่อการวิจัยจะชี้ทิศทางไว้โดยกำหนดว่าคุณลักษณะทางประชากรที่สนใจมีความเกี่ยวเนื่องถึงกัน อนึ่งเมื่อเราจำแนกกลุ่มตัวอย่างออกเป็นกลุ่มย่อย ๆ หลายกลุ่ม (cell หรือ Category) โดยทั่วไปคือ $r \times c$ หรือ $r \times c \times k$ หรือ $r \times c \times k \times l$ กลุ่ม ดังนั้น ลักษณะการแจกแจงของความถี่เหล่านี้จึงเป็นการแจกแจงแบบพหุนาม โดยกำหนดว่า p_{ij} หรือ p_{ijk} หรือ p_{ijkl} คือค่าความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จากตัวอย่าง n ชุด จะตกอยู่ใน cell ที่ (i, j) หรือ (i, j, k) หรือ (i, j, k, l) ตามลำดับ

ตารางจำแนก 2 ทาง (Two-way contingency Table : 2 CT)

2 CT คือตารางจำแนกที่จำแนกกลุ่มตัวอย่างออกเป็น 2 ทาง ตามคุณลักษณะทางประชากรที่สนใจ 2 ลักษณะ

สมมุติคุณลักษณะทางประชากร 2 ลักษณะที่สนใจ คือ A และ B และเราสามารถแบ่งคุณลักษณะ A ออกเป็น r ระดับ คือ $A_1, A_2, \ldots, A_r$ แบ่งคุณลักษณะ B ออกเป็น c ระดับ คือ $B_1, B_2, \ldots, B_c$ ให้ X_{ij} แสดงจำนวนหน่วยตัวอย่าง (ความถี่) ที่สอดคล้องกับระดับที่ i ของคุณลักษณะ A และสอดคล้องกับระดับที่ j ของคุณลักษณะ B เมื่อ $i = 1, 2, \ldots, r$ และ $j = 1, 2, \ldots, c$ [1] และความถี่ในทุก cell รวมกันจะต้องเท่ากับขนาดตัวอย่าง n หรือนัยหนึ่ง $\sum_i^r \sum_j^c X_{ij} = n$ ดังตาราง [2]

[1] หรือ X_{ij} = จำนวนความถี่ที่ตกอยู่ใน cell ที่ (i, j)

[2] p_{ij} = Probability of inclusion

คุณลักษณะ A	คุณลักษณะ B						รวม
	B_1	B_2	B_3	B_4	...	B_c	
A_1	X_{11}, p_{11}	X_{12}, p_{12}	X_{13}, p_{13}	X_{14}, p_{14}	...	$X_{1c}\ p_{1c}$	$X_{1.}, p_{1.}$
A_2	X_{21}, p_{21}	X_{22}, p_{22}	X_{23}, p_{23}	X_{24}, p_{24}	...	X_{2c}, p_{2c}	$X_{2.}, p_{2.}$
A_3	X_{31}, p_{31}	X_{32}, p_{32}	X_{33}, p_{33}	X_{34}, p_{34}	...	X_{3c}, p_{3c}	$X_{3.}, p_{3.}$
.							
.							
.							
A_4	X_{r1}, p_{r1}	X_{r2}, p_{r2}	X_{r3}, p_{r3}	X_{r4}, p_{r4}	...	X_{rc}, p_{rc}	$X_{r.}, p_{r.}$
รวม	$X_{.1}, p_{.1}$	$X_{.2}, p_{.2}$	$X_{.3}, p_{.3}$	$X_{.4}, p_{.4}$	...	$X_{.c}, p_{.c}$	$X_{..} = n$ $p_{..} = 1$

เมื่อ X_{ij} = จำนวนความถี่ที่ตกอยู่ในห้อง (cell) ที่ (i, j)

p_{ij} = Probability of Inclusion

$X_{i.} = X_{i1} + X_{i2} + \ldots + X_{ic} = \sum_{j}^{c} X_{ij}$ สำหรับทุกค่าของ i; i = 1, 2, ..., r

$X_{.j} = X_{1j} + X_{2j} + \ldots + X_{rj} = \sum_{i}^{r} X_{ij}$ สำหรับทุกค่าของ j; j = 1, 2, ..., c

$$X_{..} = \sum_{i}^{r}\sum_{j}^{c} X_{ij} = \sum_{i}^{r} X_{i.} = \sum_{j}^{c} X_{.j} = n$$

ให้ p_{ij} = ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะตกอยู่ใน cell ที่ (i, j) หรือนัยหนึ่งคือ ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะมีคุณสมบัติหรือลักษณะสอดคล้องกับระดับที่ i ของคุณลักษณะ A และระดับที่ j ของคุณลักษณะ B; i = 1, 2, ..., r; j = 1, 2, ..., c

$p_{i.} = \sum_{j}^{c} p_{ij}$ = ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะมีคุณสมบัติหรือลักษณะสอดคล้องกับระดับที่ i ของคุณลักษณะ A; i = 1, 2, ..., r

$p_{.j} = \sum_{i}^{r} p_{ij}$ = ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะมีคุณสมบัติหรือลักษณะสอดคล้องกับระดับที่ j ของคุณลักษณะ B; i = 1, 2, ..., c

$$p_{..} = \sum_i^r \sum_j^c p_{ij} = \sum_i^r p_{i.} = \sum_j^c p_{.j} = 1$$

ดังนั้น $$f(x_{11}, x_{12}, \ldots, x_{rc}; p_{11}, p_{12}, \ldots, p_{rc}) = p_{11}^{x_{11}} p_{12}^{x_{12}} \ldots p_{1c}^{x_{1c}} \ldots p_{r1}^{x_{r1}} p_{r2}^{x_{r2}} \ldots p_{rc}^{x_{rc}}$$

$$= \prod_i^r \prod_j^c p_{ij}^{x_{ij}}$$

หรือจะใช้ $$f(x_{11}, x_{12}, \ldots, x_{rc}; p_{11}, p_{12}, \ldots, p_{rc}) = \binom{n}{x_{11}\, x_{12} \ldots x_{rc}} \prod_i^r \prod_j^c p_{ij}^{x_{ij}}$$ ก็ได้

สมมุติฐานที่ต้องการทดสอบคือ คุณลักษณะ A และ B เป็นอิสระต่อกัน หรือนัยหนึ่ง เมื่อมองในรูปของความน่าจะเป็นก็คือ ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะมีคุณสมบัติสอดคล้องกับระดับที่ i ใด ๆ ของ A ได้นั้น มิจำเป็นต้องมีสาเหตุสืบเนื่องมาจากคุณสมบัติที่สอดคล้องกับระดับที่ j หนึ่งใดของ B หรือกลับกัน

นั่นคือ $Pr(A_i \mid B_j) = Pr(A_i);\ i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c$

หรือ $Pr(B_j \mid A_i) = Pr(B_j);\ i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c$

หรือ $Pr(A_i \cap B_j) = Pr(A_i)\, Pr(B_j);\ i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c$

หรือ $p_{ij} = p_{i.}\, p_{.j};\ i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c$

ดังนั้น สมมุติฐานทางสถิติที่จะทดสอบ คือ

$H_0 : p_{ij} = p_{i.} \cdot p_{.j}$ vs $H_1 : p_{ij} = p_{ij} \neq p_{i.}\, p_{.j}$ [1]

(1) การพัฒนาตัวทดสอบ

จาก H_0 และ H_1 เราสามารถแจง parameter space ออกมาได้ดังนี้

$$\omega = \{p_{i.}, p_{.j};\ 0 \leqslant p_{i.} \leqslant 1,\ 0 \leqslant p_{.j} \leqslant 1;\ i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c\}$$

และ

$$\Omega = \{p_{ij};\ 0 \leqslant p_{ij} \leqslant 1;\ i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c\}$$

---

[1] H_0 : ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะตกอยู่ใน cell ที่ (i, j) มีค่าเท่ากับความน่าจะเป็นที่หน่วยตัวอย่างจะตกอยู่ในระดับที่ i ของ A และ ระดับที่ j ของ B หรือ $Pr(A_i \cap B_j) = P(A_i)\,P(B_j)$ หรือความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะมีคุณสมบัติสอดคล้องกับระดับที่ i ของคุณลักษณะ A ได้นั้น มิจำเป็นต้องมีสาเหตุมาจากการมีคุณสมบัติที่สอดคล้องกับระดับที่ j หนึ่งใดของ B

H_1 : ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ ที่จะมีคุณสมบัติสอดคล้องกับระดับที่ i ของ A ได้นั้น จำเป็นต้องมีสาเหตุพัวพัน (Interaction) มาจากระดับที่ j ของคุณลักษณะ B

ดังนั้น $L_\omega = \prod_i^r \prod_j^c (p_{i.}p_{.j})^{X_{ij}}$ และ $L_\Omega = \prod_i^r \prod_j^c (p_{ij})^{X_{ij}}$

พิจารณา L_ω จะพบว่า

$$L_\omega = \prod_i^r \prod_j^c (p_{i.}p_{.j})^{X_{ij}} = \prod_i^r \prod_j^c (p_{i.})^{X_{ij}} \prod_i^r \prod_j^c (p_{.j})^{X_{ij}}\ ^{1}$$

$$= \prod_i^r \{(p_{i.})^{X_{i1}} (p_{i.})^{X_{i2}} \ldots (p_{i.})^{X_{ic}}\} \cdot \prod_j^c \{(p_{.j})^{X_{1j}} (p_{.j})^{X_{2j}} \ldots (p_{.j})^{X_{rj}}\}$$

$$= \prod_i^r (p_{i.})^{\sum_j^c X_{ij}} \prod_j^c (p_{.j})^{\sum_i^r X_{ij}}$$

$$=> L_\omega = \prod_i^r (p_{i.})^{X_{i.}} \prod_j^c (p_{.j})^{X_{.j}}$$

เนื่องจากเซท ω ประกอบด้วยพารามิเตอร์ที่ไม่ทราบค่า $(r-1)+(c-1)$ ตัวคือ $p_{1.}$, $p_{2.}, \ldots, p_{(r-1).}$ และ $p_{.1}, p_{.2}, \ldots, p_{.(c-1)}$ ทั้งนี้ เพราะ

$$p_{r.} = 1 - \sum_i^{r-1} p_{i.} = 1 - (p_{1.} + p_{2.} + \ldots + p_{(r-1).}) \text{ และ } p_{.c} = 1 - \sum_j^{c-1} p_{.j}$$

$$= 1 - (p_{.1} + p_{.2} + \ldots + p_{.(c-1)})$$

ดังนั้น เราจึงประมาณค่าพารามิเตอร์เพียง $r + c - 2$ ตัว ในที่นี้จะเริ่มด้วยการประมาณค่า $p_{i.}; i = 1, 2, \ldots, r-1$ ก่อน แล้วจึงค่อยประมาณค่า $p_{.j}; j = 1, 2, \ldots, c-1$

จาก $L_\omega = \prod_i^r (p_{i.})^{X_{i.}} \prod_j^c (p_{.j})^{X_{.j}}$

แทนที่ $p_{r.}$ ด้วย $1 - \sum_i^{r-1} p_{i.}$ และแทนที่ $p_{.j}$ ด้วย $1 - \sum_j^{c-1} p_{.j}$

$$=> L_\omega = (p_{1.})^{X_{1.}} (p_{2.})^{X_{2.}} \ldots (p_{(r-1).})^{X_{(r-1).}} (1 - \sum_i^{r-1} p_{i.})^{X_{r.}} (p_{.1})^{X_{.1}} (p_{.2})^{X_{.2}} \ldots p_{.(c-1)}^{X_{.(c-1)}} (1 - \sum_j^{c-1} p_{.j})^{X}$$

---

[1] เพื่อแสดงให้เห็นความจริงของ identity นี้ ขอให้นักศึกษาทดลองกระจายกับตัวอย่างขนาดเล็ก เช่น $\prod_i^2 \prod_j^2 (p_{i.}\ p_{.j})^{X_{ij}}$ โดยไล่เรียง subscript ทีละตัว อาจเริ่มจาก j ก่อน แล้วเริ่ม i ภายหลังก็ได้

$$\frac{\partial}{\partial p_i} \ln L_\omega = 0; i = 1, 2, \ldots, r-1$$

ดังนั้น $0 + 0 + \ldots + 0 + \dfrac{X_{i.}}{p_{i.}} + 0 + \ldots + \dfrac{X_{r.}(-1)}{(1 - \sum_{i}^{r-1} p_{i.})} = 0$

$; i = 1, 2, \ldots, r - c$

$$=> \frac{X_{i.}}{p_{i.}} = \frac{X_{r.}}{p_{r.}}; i = 1, 2, \ldots, r-1$$

$$p_{i.} = (p_{r.}/X_{r.}) X_{i.}; i = 1, 2, \ldots, r-1 \qquad \ldots\ldots\ldots.(1)$$

รวมตลอดในทุกค่าของ i ตั้งแต่ 1 ถึง r

$$\sum_{i}^{r} p_{i.} = (p_{r.}/X_{r.}) \sum_{i}^{r} X_{i.}$$

$$1 = (p_{r.}/X_{r.})n$$

ดังนั้น $p_{r.}/X_{r.} = 1/n$ ....(2)

แทนค่า (2) ใน (1) $=> \hat{p}_{i.} = X_{i.}/n; i = 1, 2, \ldots, r-1$

และ $\hat{p}_{r.} = 1 - (\hat{p}_{1.} + \hat{p}_{2.} + \ldots + \hat{p}_{(r-1).})$

$$= 1 - (\frac{X_{1.}}{n} + \frac{X_{2.}}{n} + \ldots + \frac{X_{(r-1).}}{n})$$

$$= \frac{X_{r.}}{n}$$

ในทำนองเดียวกัน เมื่อดิฟเฟอร์เรนชิเอท $\ln L_\omega$ เทียบต่อ $p_{.j}$ เราสามารถพิสูจน์ได้ว่า

$p_{.j} = \dfrac{X_{.j}}{n}; j = 1, 2, \ldots, c-1$ และ $\hat{p}_{.c} = \dfrac{X_{.c}}{n}$

แทนค่า $\hat{p}_{i.}, \hat{p}_{.j}$ ใน L_ω จะได้ $L_{\hat{\omega}}$ ดังนี้

$$L_{\hat{\omega}} = \prod_i^r \left(\frac{X_{i.}}{n}\right)^{X_{i.}} \prod_j^c \left(\frac{X_{.j}}{n}\right)^{X_{.j}}$$

$$= \frac{\prod_i^r (X_{i.})^{X_{i.}}}{n^{X_{1.}} n^{X_{2.}} \ldots n^{X_{r.}}} \quad \frac{\prod_j^c (X_{.j})^{X_{.j}}}{n^{X_{.1}} n^{X_{.2}} \ldots n^{X_{.c}}}$$

$$= \frac{1}{n^n} \prod_i^r (X_{i.})^{X_{i.}} \frac{1}{n^n} \prod_j^c (X_{.j})^{X_{.j}}$$

พิจารณา $L_\Omega = \prod_i^r \prod_j^c (p_{ij})^{X_{ij}}$

$$= \prod_i^{r-1} \prod_j^{c-1} (p_{ij})^{X_{ij}} \prod_i^{r-1} (p_{ic})^{X_{ic}} \prod_j^{c-1} (p_{rj})^{X_{rj}}$$

$$\{1 - \sum_i^{r-1} \sum_j^{c-1} p_{ij} - \sum_i^{r-1} p_{ic} - \sum_j^{c-1} p_{rj}\}^{X_{rc}}$$ [1]

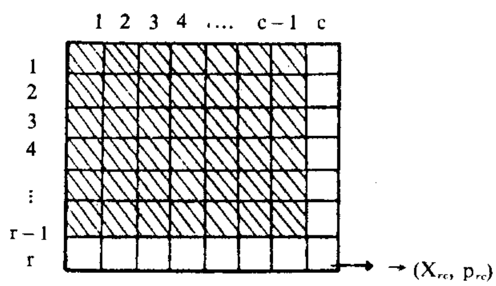
$$\frac{\partial}{\partial p_{ij}} \ln L_\Omega = 0; i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$$ ยกเว้น p_{rc}

$$\Rightarrow 0 + 0 + \ldots + 0 + \frac{X_{ij}}{p_{ij}} + 0 + \ldots + 0 + \frac{(-1)\,X_{rc}}{1 - \sum_i^{r-1} \sum_j^{c-1} p_{ij} - \sum_i^{r-1} p_{ic} - \sum_j^{c-1} p_{rj}}$$

$i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$ ยกเว้น p_{rc}

---

[1] ดูไดอะแกรม



$$=> \frac{X_{ij}}{p_{ij}} = \frac{X_{rc}}{p_{rc}}; i = 1, 2, ..., r; j = 1, 2, ..., c$$

$$p_{ij} = \frac{p_{rc}}{X_{rc}} \cdot X_{ij}; i = 1, 2, ..., r; j = 1, 2, ..., c \qquad ....(1)$$

$$\sum_i^r \sum_j^c p_{ij} = \frac{p_{rc}}{X_{rc}} \sum_i^r \sum_j^c X_{ij}$$

$$1 = \frac{p_{rc}}{X_{rc}} \cdot n$$

$$\frac{p_{rc}}{X_{rc}} = \frac{1}{n} \qquad ....(2)$$

แทน (2) ใน (1)

$$=> \hat{p}_{ij} = \frac{X_{ij}}{n}; i = 1, 2, ..., r; j = 1, 2, ..., c \text{ ยกเว้น } p_{rc}$$

และเราสามารถพิสูจน์ได้ว่า $\hat{p}_{rc} = \frac{X_{rc}}{n}$

ดังนั้น $L_{\hat{\Omega}} = \prod_i^r \prod_j^c (\frac{X_{ij}}{n})^{x_{ij}} = \frac{1}{n^{\{\Sigma\Sigma X_{ij}\}}} \prod_i^r \prod_j^c X_{ij}^{x_{ij}}$

$$=> L_{\hat{\Omega}} = \frac{1}{n^n} \prod_i^r \prod_j^c (X_{ij})^{x_{ij}}$$

โดยอาศัย MLR เราจะปฏิเสธสมมุติฐานหลักเมื่อ $\lambda = \frac{L_{\hat{\omega}}}{L_{\hat{\Omega}}} \leqslant k$

$$\lambda = \{\frac{1}{n^n} \cdot \frac{1}{n^n} \prod_i^r (X_{i.})^{x_{i.}} \prod_j^c (X_{.j})^{x_{.j}}\} / \{\frac{1}{n^n} \prod_i^r \prod_j^c (X_{ij})^{x_{ij}}\} \leqslant k$$

$$=> \lambda = \frac{\prod_i^r (X_{i.})^{x_{i.}} \prod_j^c (X_{.j})^{x_{.j}}}{n^n \prod_i^r \prod_j^c (X_{ij})^{x_{ij}}} \leqslant k$$

นั่นคือ ปฏิเสธสมมุติฐานหลัก ณ ระดับนัยสำคัญ α เมื่อ $\lambda \leqslant k$ โดยที่ค่าของ k สามารถคำนวณได้จากสมการต่อไปนี้คือ

$$\int_{-\infty}^{k} f_{\Lambda}(\lambda)\,dx = \alpha^{1}$$

อย่างไรก็ตาม pd f ของ λ หาได้ค่อนข้างยากและซับซ้อนมาก[2] ดังนั้นในทางปฏิบัติจึงจำเป็นต้องประมาณ pd f ของ λ โดยอาศัยทฤษฎี 7.1 ทั้งนี้มีข้อจำกัดว่าขนาดตัวอย่าง n ต้องใหญ่มากพอ $(n \to \infty)$

$$=> \; -2\ln\lambda \sim \chi^2$$

สำหรับ df ของ χ^2 ให้พิจารณา Reparameterized Space "กล่าวคือจาก Ω จะพบว่าเสนอพารามิเตอร์ไว้เพียง $rc-1$ ค่า ค่าสุดท้ายคือ p_{rc} ไม่จำเป็นต้องกำหนดเพราะ $\sum_{i}^{r}\sum_{j}^{c} p_{ij} = 1$ สำหรับใน ω จะพบว่าเสนอพารามิเตอร์ไว้เพียง $(r-1)+(c-1)$ ค่าคือ $p_{i.}\,;\, i = 1, 2, \ldots, r-1$ และ $p_{.j}\,;\, j = 1, 2, \ldots, c-1$ ค่าสุดท้ายคือ $p_{r.}$ และ $p_{.c}$ ไม่จำเป็นต้องกำหนดเพราะ $\sum_{i}^{r} p_{i.} = 1$ และ $\sum_{j}^{c} p_{.j} = 1$

ดังนั้น Reparemeterized Space จึงประกอบด้วยพารามิเตอร์เพียง $(rc-1)-(r+c-2)$ $= rc-r-c+1 = (r-1)(c-1)$ ค่า

ดังนั้น โดยอาศัยทฤษฎี 7.1 $-2\ln\lambda \sim \chi^2_{(r-1)(c-1)}$ เมื่อ

$$\lambda = \{\prod^{r} (X_{i.})^{x_{i.}} \prod^{c} (X_{.j})^{x_{.j}}\} / \{n^n \prod_{i}^{r}\prod_{j}^{c} (X_{ij})^{x_{ij}}\}$$

นั่นคือ ปฏิเสธสมมุติฐานหลักเมื่อ $-2\ln\lambda = 2\ln\left\{\dfrac{\Pi(X_{i.})^{x_{i.}} \cdot \Pi(X_{.j})^{x_{.j}}}{n^n\,\Pi\,\Pi(X_{ij})^{x_{ij}}}\right\} > -2\ln k = k'$

และโดยอาศัยเทคนิคทำนองเดียวกับตอน 6.3.2 เราสามารถพิสูจน์ได้ว่า เมื่อ $n \to \infty$ แล้ว เราจะปฏิเสธสมมุติฐานหลักเมื่อ

---

[1] ในภาษากรีก Λ คืออักษรตัวพิมพ์ขณะที่ λ คืออักษรตัวเขียน อ่านว่าแลมด้า

[2] อ่านเพิ่มเติมจาก Mood, Graybill and Boe, Introduction to the Theory of Statistics, 3rd ed.. p.455-9

[3] ทฤษฎี 7.1