Journal of educational equipment: Applied research, Volume 2, Issue 305 (January 2024)
ISSN 1859 - 0810

Dạy học khoảng tin cậy cho tỷ lệ với sự hỗ trợ của phần mềm STATA

Đào Hồng Nam

TS. Trường Đại học Y Dược TP Hồ Chí Minh

Received: 5/12/2023; Accepted: 8/12/2023; Published: 11/12/2023

Abstract: *The confidence interval for the ratio p is one of the mandatory and necessary knowledge in the curriculum of the Statistics Probability module at the university level. There are many methods for estimating CIs for different rates, but the CI most used in statistical probability textbooks in Vietnam that we refer to is the Wald CI because of its simplicity in calculation and application in practice. international. However, the conditions for use of the Wald CI must be satisfied, otherwise the estimated CI for the population will be inaccurate and even have no use value such as CI containing negative values or exceeding 1.*

This article presents some alternative methods in teaching when the conditions of the Wald CI are not satisfied through an illustrative example with the support of Stata 17 software.

Keywords*: Teaching, confidence interval, rate, accuracy, testing.*

1. Đặt vấn đề

Khoảng tin cậy (KTC) cho tỷ lệ nhị thức p được sử dụng rất phổ biến, nhất là KTC Wald (Brown et al., 2001, 2002). KTC Wald rất dễ tính toán, ngay cả khi thực hiện thủ công không có sự trợ giúp của các phần mềm thống kê, so với những KTC khác như Agresti-Coull (1998), Clopper-Pearson (1934), Wilson (1927), … do đó, phần lớn các giáo trình đại học chuyên ngành khoa học sức khỏe ở Việt Nam đều chỉ trình bày KTC Wald mà không có những KTC khác thay thế khi những điều kiện của KTC Wald bị vi phạm. Một số tác giả cũng đã chỉ ra nhược điểm của KTC Wald và đề xuất thay thế bằng các khoảng khác như Per Gösta Andersson (2023), Brown, L., Cai, T. and DasGupta, A. (2001, 2002).

Điều này có thể dẫn đến những kết luận sai lầm của sinh viên (SV) trong quá trình học tri thức này do KTC vượt ra ngoài đoạn [0; 1]. Một trong những sai lầm sẽ xảy ra khi tỷ lệ gần bằng 0 (như khi nghiên cứu về các bệnh hiếm) hoặc gần bằng 1 (khi đánh giá độ chính xác của một xét nghiệm: độ nhạy, độ đặc hiệu).

2. Nội dung nghiên cứu

2.1.KTC Wald

KTC cho tỷ lệ p dựa trên phân phối xấp xỉ chuẩn được giới thiệu phổ biến trong các giáo trình đại học ở Việt Nam là KTC tiêu chuẩn (KTC Wald).

Ước lượng KTC của tỷ lệ tổng thể p, thông qua tỷ lệ mẫu $\hat{p} = \frac{k}{n}$ trong đó k là số phần tử có đặc tính X cần ước lượng trong n phần tử của mẫu, là bài toán xác định khoảng (p_1, p_2) sao cho $P(p_1 \le p \le p_2) = 1 - \alpha$, với $(1 - \alpha) = \gamma$ là độ tin cậy cho trước.

Xét biến ngẫu nhiên $Y \sim B(n; p)$ có phân phối nhị thức với $\mu = E(Y) = n\hat{p}; \sigma^2 = D(Y) = n\hat{p}(1-\hat{p})$.

Người ta chứng minh được rằng khi n khá lớn sao cho $nmin\{\hat{p}, 1-\hat{p}\} \ge 10$ thì tỷ lệ mẫu $\hat{p}$ sẽ có phân phối xấp xỉ PPC với trung bình $E(\hat{p}) = p$ và phương sai $D(\hat{p}) = \frac{p(1-p)}{n}$ tức là $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Vậy với biến ngẫu nhiên $Z \sim N(0,1)$ thì $P(-z_{1-\alpha/2} \le Z \le z_{1-\alpha}) = 1-\alpha$.

Do $\hat{p} \sim N\left(p, p(1-p)/n\right)$ nên $Z = \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0,1)$

Khi đó:

$$P\left(-z_{1-\frac{\alpha}{2}} \le Z \le z_{1-\alpha}\right) = 1-\alpha \Leftrightarrow P\left(-z_{1-\frac{\alpha}{2}} \le \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{p(1-p)}} \le z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

Vậy với độ tin cậy $(\gamma = 1-\alpha)$ cho trước, tỷ lệ tổng thể p được xác định

$$\hat{p} - C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Với $p_1 = \hat{p} - C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $p_2 = \hat{p} + C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Thông thường KTC cho tỷ lệ p được viết là:

$$CI_s = \hat{p} \pm C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} (1)$$

Với $C = z_{1-\alpha/2}$ là giá trị thỏa mãn $P(|Z| > C) = 1-\alpha/2$ và Z là biến ngẫu nhiên có phân phối chuẩn tắc, ký hiệu là $Z \sim N(0;1)$.

KTC trong (1) được gọi là KTC Wald vì nó xuất phát từ thử nghiệm mẫu lớn Wald cho trường hợp nhị thức. KTC Wald rất dễ tính toán ngay cả khi tính toán thủ công. Mức độ phổ biến của nó trong thống kê hầu như không có tri thức nào có thể so sánh được.

Tuy nhiên, người ta thừa nhận rộng rãi rằng xác suất bao phủ thực tế của KTC Wald là kém khi p gần 0 hoặc 1. Trong các giáo trình giảng dạy đại học, KTC Wald thường được trình bày kèm theo lời cảnh báo rằng nó chỉ nên được sử dụng khi $n.min\, p(1-p) \geq 5$ (hoặc 10) (*)

Ví dụ: Xét nghiệm T được thực hiện trên 110 người trong quần thể D có 3 người dương tính.

Tìm KTC 95% cho tỷ lệ dương tính của T.

Nếu không chú ý đến điều kiện (*) mà thực hiện ngay KTC Wald sẽ có kết quả:

$$\hat{p} = \frac{3}{110} = 0.027$$

$$CI_{Wald} = \hat{p} \pm C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.027 \pm 1.96\sqrt{\frac{0.027(1-0.027)}{110}}$$

$$= 0.027 \pm 0.03 = [-0.003; 0.057]$$

Rõ ràng đây là một kết quả sai vì KTC cho tỷ lệ chứa khoảng âm.

Để khắc phục nhược điểm này, nhất là khi tỷ lệ $\hat{p}$ rất gần 0 hoặc rất gần 1 thì có thể sử dụng các KTC được đề xuất bao gồm Clopper–Pearson, Wilson và Agresti-Coull.

2.2. KTC Agresti-Coull

KTC này cũng khá dễ tính toán và có hình thức rất giống KTC Wald. Công thức tính KTC là Agresti-Coull là:

$$CI_{AC} = \tilde{p} \pm C\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} (2)$$

Với $\tilde{n} = n+4; \tilde{p} = \frac{X+2}{n+4}$ và $C = z_{1-\alpha/2}$ là giá trị thỏa mãn $P(|Z| > C) = 1-\alpha/2$ và Z là biến ngẫu nhiên có phân phối chuẩn tắc, ký hiệu là $Z \sim N(0;1)$.

Trong ví dụ trên:

$$\tilde{n} = n+4 = 110+4 = 114; \tilde{p} = \frac{X+2}{n+4} = \frac{3+2}{110+4} = \frac{5}{114}$$

Thay vào công thức (2) ta có KTC Agresti-Coull là:

$$CI_{AC} = \tilde{p} \pm C\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \frac{5}{114} \pm 1.96\sqrt{\frac{\frac{5}{114}\left(1-\frac{5}{114}\right)}{114}}$$

$$= 0.044 \pm 0.038 = [0.006; 0.082]$$

2.3. KTC Wilson

KTC này do Edwin Bidwell Wilson đề xuất vào năm 1927. Khác với KTC tiêu chuẩn, KTC Wilson là KTC không đối xứng. KTC này được sử dụng với một số ít lần thử nghiệm ($n \leq 40$) và/hoặc tỷ lệ của biến cố cần quan tâm là những giá trị rất gần 0 hoặc 1.

KTC Wilson có dạng:

$$CI_W = \frac{2k+C^2}{2(n+C^2)} \pm \frac{C}{n+C^2}\sqrt{n\hat{p}(1-\hat{p}) + \frac{C^2}{4}} (3)$$

Trong ví dụ

$$CI_W = \frac{2\times 3+1.96^2}{2(110+1.96^2)} \pm \frac{1.96}{110+1.96^2}\sqrt{110\times\frac{3}{110}\left(1-\frac{3}{110}\right)+\frac{1.96^2}{4}}$$

$$= [0.009; 0.077]$$

2.4. KTC Clopper-Pearson

$$\left(1+F_{\frac{\alpha}{2}}(\vartheta_1,\vartheta_2)\times\frac{\hat{q}+\frac{1}{n}}{\hat{p}}\right)^{-1} \leq p \leq \left(1+\frac{\hat{q}}{F_{\frac{\alpha}{2}}(\vartheta_3,\vartheta_4)\left(\frac{1}{n}+\hat{p}\right)}\right)^{-1} (4)$$

$F_{\frac{\alpha}{2}}$ là giá trị ngưỡng trong phân phối Fisher với các bậc tự được tính như sau:

$$\vartheta_1 = 2(n\hat{q}+1); \vartheta_2 = 2n\hat{p}; \vartheta_3 = 2(n\hat{p}+1); \vartheta_4 = 2n\hat{q}$$

Trong ví dụ

$$\hat{p} = \frac{3}{110} \Rightarrow \hat{q} = 1-\hat{p} = \frac{107}{110}$$

$$\vartheta_1 = 2(n\hat{q}+1) = 2\left(110\times\frac{107}{110}+1\right) = 216; \vartheta_2 = 2n\hat{p} = 2\times 110\times\frac{3}{110} = 6$$

$$\Rightarrow F_{\frac{\alpha}{2}}(\vartheta_1,\vartheta_2) = F_{0.025}(216;6) = 4.88$$

$$\vartheta_3 = 2(n\hat{p}+1) = 2\left(110\times\frac{3}{110}+1\right) = 8; \vartheta_4 = 2n\hat{q} = 2\times 110\times\frac{107}{110} = 214$$

$$\Rightarrow F_{\frac{\alpha}{2}}(\vartheta_3,\vartheta_4) = F_{0.025}(8;214) = 2.252$$

Thay vào công thức (4):

$$\left(1+4.88\times\frac{\frac{107}{110}+\frac{1}{110}}{\frac{3}{110}}\right)^{-1} \leq p \leq \left(1+\frac{\frac{107}{110}}{2.252\left(\frac{1}{110}+\frac{3}{103}\right)}\right)^{-1}$$

$$0.006 \leq p \leq 0.078$$

Sau đây chúng tôi tóm tắt kết quả tính toán KTC bằng các phương pháp như đã trình bày ở trên:

Khoảng	Số lần thành công	Cỡ mẫu	Tỷ lệ mẫu	Cận dưới	Cận trên
Wald	3	110	0.027	- 0.003	0.058
Clopper-Pearson exact	3	110	0.027	0.006	0.078
Wilson	3	110	0.027	0.009	0.077
Agresti-Coull	3	110	0.027	0.006	0.081

Khi dạy học trên lớp và sử dụng trong phòng thi, tính đơn giản của KTC trong quá trình tính toán cũng cần phải xem xét. Đặc biệt là trong phòng thi, nếu SV chỉ được phép sử dụng máy tính cầm tay (không sử dụng phần mềm hỗ trợ) thì tính đơn giản càng phải được đặt lên hàng đầu. Xem xét các yếu tố này, một số tác giả (Brown, L. D., Cai, T., & Dasgupta, A, 2001) khuyến nghị rằng đối với $n \leq 40$ thì nên sử dụng khoảng Wilson. Đối với $n > 40$, khoảng Wilson và Agresti–Coull đều rất giống nhau, và do đó dạng đơn giản nhất là khoảng Agresti–Coull sẽ là lựa chọn tốt. Ngay cả đối với cỡ mẫu nhỏ hơn, khoảng Agresti–Coull vẫn được ưa chuộng hơn so với khoảng tiêu chuẩn.

Để khắc phục những khó khăn khi phải thực hiện ước lượng KTC Clopper-Pearson, Wilson hoặc Agresti–Coull. Giảng viên có thể hướng dẫn sinh viên sử dụng phần mềm Stata 17 để thực hành tính toán. Cụ thể là các câu lệnh sau đây:

Để tìm KTC 95% cho tỷ lệ p, sử dụng lệnh: cii prop 110 3, wald

Kết quả như hình 2.1.

Variable	Obs	Proportion	Std. err.	Binomial Wald [95% conf.	interval]
	110	.0272727	.0155297	0	.0577104

The Wald interval was clipped at the lower endpoint.

Hình 2.1. KTC tiêu chuẩn (khoảng Wald) trong Stata 17

Do KTC cho tỷ lệ chứa khoảng âm nên Stata có cảnh báo về việc cắt bỏ giới hạn dưới của khoảng này.

Nếu sử dụng KTC chính xác (Cloper-Pearson) thì sử dụng câu lệnh: cii prop 110 3, exact

Kết quả như hình 2.2.

Variable	Obs	Proportion	Std. err.	Binomial exact [95% conf.	interval]
	110	.0272727	.0155297	.00566	.0776368

Hình 2.2. KTC chính xác (khoảng Cloper-Pearson) trong Stata

Nếu sử dụng KTC Wilson thì câu lệnh là: cii prop 110 3, wilson

Kết quả của KTC Wilson khá giống với kết quả của KTC Cloper-Pearson

Variable	Obs	Proportion	Std. err.	Wilson [95% conf.	interval]
	110	.0272727	.0155297	.009318	.0771308

Hình 2.3. KTC Wilson trong Stata

Nếu sử dụng KTC Agresti-Coull thì câu lệnh là: cii prop 110 3, agresti

Variable	Obs	Proportion	Std. err.	Agresti-Coull [95% conf.	interval]
	110	.0272727	.0155297	.0058678	.080581

Hình 2.4. KTC Agresti-Coull trong Stata 17

3. Kết luận

Ước lượng KTC cho tỷ lệ là một trong những nội dung trong chương trình giảng dạy xác suất thống kê ở bậc đại học tại Việt Nam. Trong các giáo trình này, các tác giả thường chỉ trình bày một phương pháp ước lượng KTC tiêu chuẩn (KTC Wald) do việc tính toán KTC này khá nhanh chóng và dễ dàng ngay cả khi thực hiện thủ công hoặc phương tiện máy tính cầm tay đơn giản. Tuy nhiên, khi điều kiện của KTC Wald không được thỏa mãn thì các KTC tìm được có thể không hợp lệ như tình huống trong bài báo này đã chỉ ra. Để khắc phục nhược điểm của KTC Wald, chúng tôi đề xuất sử dụng một số KTC thay thế như Clopper – Pearson, Wilson, Agresti–Coull. Trong các khoảng thay thế này, khoảng Agresti–Coull là đơn giản và dễ thực hiện nhất nên chúng tôi đề xuất đưa KTC Agresti–Coull vào trong các giáo trình thống kê cho sinh viên học tập, nghiên cứu nhằm đơn giản hóa quá trình tính toán và phù hợp với phương pháp học và thi hiện nay khi sinh viên chỉ được phép sử dụng máy tính cầm tay trong giờ thi có giới hạn. Trong các bài giảng trên lớp hoặc trong các nghiên cứu khoa học, giảng viên có thể hướng dẫn sinh viên sử dụng phần mềm Stata 17 để hỗ trợ tính toán các KTC này.

Ghi chú: Đề tài này được nhận kinh phí tài trợ từ Trường Đại học Y Dược TP Hồ Chí Minh.

Tài liệu tham khảo

1.Agresti, A., & Coull, B. A. (1998). *Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician*, *52*(2), 119–126.

2.Brown, L., Cai, T. and DasGupta, A. (2001). *Interval estimation for a binomial proportion. Statistical Science 16, 101–117.*

3.Brown, L., Cai, T. and DasGupta, A. (2002). *Confidence intervals for a binomial proportion and asymptotic expansions. The Annals of Statistics 30, 160–201.*

4.Clopper, C. and Pearson, E. (1934). *The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413.*