Buletinul Ştiinţific al Universităţii “Politehnica” din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII

TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 50(64), Fascicola 2, 2005

On the coverage probability of the Clopper-Pearson confidence interval

Dominique Pastor [1]

***Abstract*— Given $\alpha \in (0,1)$, the coverage probability of the $100(1-\alpha)\%$ Clopper-Pearson Confidence Interval (CPCI) for estimating a binomial parameter p is proved to be larger than or equal to $1-\alpha/2$ for sample sizes less than a bound that depends on p. This is a mathematical evidence that, as noticed in recent papers on the basis of numerical results, the CPCI coverage probability can be much higher than the desired confidence level and thence, that the Clopper-Pearson approach is mostly inappropriate for forming confidence intervals with coverage probabilities close to the desired confidence level.**

I. Introduction

INTERVAL estimation of a binomial proportion p is one of the most basic and methodologically important problems in practical statistics. A considerable literature exists on the topic and manifold methods have been suggested for bracketing a binomial proportion within a confidence interval.

Among those methods, the Clopper-Pearson Confidence Interval (CPCI), originally introduced in 1934 by Clopper and Pearson [4], is often referred as the ”exact” procedure by some authors for it derives from the binomial distribution without resorting to any approximation to the binomial. However, despite this ”exacteness”, the CPCI is basically conservative because of the discreteness of the binomial distribution: given $\alpha \in (0,1)$, the coverage probability of the $100(1-\alpha)\%$ CPCI is above or equal to the nominal confidence level $(1-\alpha)$.

In this paper, it is proved that this coverage probability is actually larger than or equal to $1-\alpha/2$ if the sample size is less than $\ln(\alpha/2)/\ln(\max(p,1-p))$. This bound basically depends on the proportion itself. Thence, as suggested by numerical results presented in [2] and [3], the CPCI is not suitable in practical situations when getting a coverage probability close to the specified confidence level is more desirable than guaranteeing a coverage probability above or equal to the said confidence level.

II. Theoretical results

Throughout the rest of this paper, α stands for some real value in the interval $(0,1)$ and n for some natural number.

We start by giving a description of the $100(1-\alpha)\%$ CPCI, that is the CPCI with level of confidence $(1-\alpha)$ where $\alpha \in (0,1)$. This description is purposely brief for it focuses only on the material needed for stating proposition 1 below. For further details on the construction of the CPCI, the reader can refer to numerous papers and textbooks on statistics ([2], [3], [5], [6] amongst many others).

Let $(\ell_k)_{k\in\{0,\ldots,n\}}$ be the sequence defined as follows. We put

$$\ell_0 := 0. \tag{1}$$

For $k \in \{1,\ldots,n\}$, ℓ_k is defined as the unique solution in $(0,1)$ for θ in the equation

$$\sum_{i=k}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} = \alpha/2. \tag{2}$$

Similarly, the sequence $(u_k)_{k\in\{0,\ldots,n\}}$ is defined as follows. We set

$$u_n := 1 \tag{3}$$

and, for $k \in \{0,\ldots,n-1\}$, u_k is the unique solution in $(0,1)$ for θ in the equation

$$\sum_{i=0}^{k} \binom{n}{i} \theta^i (1-\theta)^{n-i} = \alpha/2. \tag{4}$$

Let X henceforth stand for a binomial variate for sample size n and binomial parameter $p \in [0,1]$. In other words, X stands for the total number of successes in n independent trials with constant probability p of success.

Since X is valued in $\{0,1,\ldots,n\}$, the random variables ℓ_X and u_X are well-defined. The $100(1-\alpha)\%$ CPCI for size n is then the random interval (ℓ_X, u_X) whose lower and upper endpoints are ℓ_X and u_X respectively. The CPCI coverage probability is defined as the probability $P(\{\ell_X < p < u_X\})$ that the CPCI actually brackets the true value p of the proportion. The $100(1-\alpha)\%$ CPCI is known to be conservative in the sense that its coverage probability is always above or equal to the confidence level ([2], [3], [4]). This standard result is refined by the following proposition whose proof is given in section 3.

[1] ENST Bretagne, Technopôle de Brest-Iroise, CS - 83818, 29238 Brest Cedex 3, FRANCE, e-mail: dominique.pastor@enst-bretagne.fr

Proposition 2.1: With the same notations as above,

(i) *If* $n < -\ln(\alpha/2)/\ln(2)$*, then*

$$\begin{array}{rclll} P\left(\{\ell_X < p < u_X\}\right) & = & 0 & \textit{for} & p = 0, \\ & \geq & 1 - \alpha/2 & \textit{for} & 0 < p \leq (\alpha/2)^{1/n}\,, \\ & = & 1 & \textit{for} & (\alpha/2)^{1/n} < p < 1 - (\alpha/2)^{1/n}\,, \\ & \geq & 1 - \alpha/2 & \textit{for} & 1 - (\alpha/2)^{1/n} \leq p < 1, \\ & = & 0 & \textit{for} & p = 1. \end{array}$$

(ii) *If* $n \geq -\ln(\alpha/2)/\ln(2)$*, then*

$$\begin{array}{rclll} P\left(\{\ell_X < p < u_X\}\right) & = & 0 & \textit{for} & p = 0, \\ & \geq & 1 - \alpha/2 & \textit{for} & 0 < p < 1 - (\alpha/2)^{1/n}, \\ & \geq & 1 - \alpha & \textit{for} & 1 - (\alpha/2)^{1/n} \leq p \leq (\alpha/2)^{1/n}, \\ & \geq & 1 - \alpha/2 & \textit{for} & (\alpha/2)^{1/n} < p < 1, \\ & = & 0 & \textit{for} & p = 1. \end{array}$$

The following result derives from the proposition above.

Lemma 2.2: With the same notations as above, for every proportion $p \in (0, 1)$ *and every natural number* n *less than* $\ln(\alpha/2)/\ln(\max(p, 1-p))$*, the CPCI coverage probability is larger than or equal to* $1 - \alpha/2$*.*

PROOF: Since $\max(p, 1-p)$ is larger than or equal to $1/2$, $-\ln(\alpha/2)/\ln(2)$ is less than $\ln(\alpha/2)/\ln(\max(p, 1-p))$. If $n < -\ln(\alpha/2)/\ln(2)$, the result then follows from proposition 1, statement (i). If $-\ln(\alpha/2)/\ln(2) \leq n < \ln(\alpha/2)/\ln(\max(p, 1-p))$, p is either larger than $(\alpha/2)^{1/n}$ or smaller than $1 - (\alpha/2)^{1/n}$ and the result follows from proposition 1, statement (ii).

■

The theoretical results above are thus mathematical evidences that, as suggested in [2] and [3], the CPCI is inaccurate in the sense that "*[...]its actual coverage probability can be much larger than* $1-\alpha$ *unless the sample size* n *is quite large.*" ([3, Sec. 4.2.1]).

Figures 1, 2 and 3 illustrate the foregoing by displaying the coverage probability of the 95% CPCI for $p = 0.1$, $p = 0.05$ and $p = 0.01$ respectively and sample sizes ranging from 1 to 500. In each figure, the value of $\ln(\alpha/2)/\ln(\max(p, 1-p))$ is represented by a vertical red line. On the left hand side of this line, coverage probabilities are all larger than or equal to 97.5%; on the right hand side of this same line, coverage probabilities can be less than 97.5% and even close to 95%.

III. PROOF OF PROPOSITION 2.1

Given $\theta \in [0, 1]$, let P_θ stand for the distribution of a binomial variate for sample size n with binomial parameter θ and let F_θ be the distribution function defined for every real value x by $F_\theta(x) = P_\theta\left((-\infty, x]\right)$. It is then convenient to set $G_\theta(x) = P_\theta\left([x, \infty)\right)$ for every real value x. According to the definitions of F_θ and G_θ, the left hand sides in (2) and (4) are equal to $G_\theta(k)$ and $F_\theta(k)$ respectively. Therefore, by definition of ℓ_k and u_k,

$$G_{\ell_k}(k) = \alpha/2 \text{ for every } k \in \{1, \ldots, n\} \qquad (5)$$

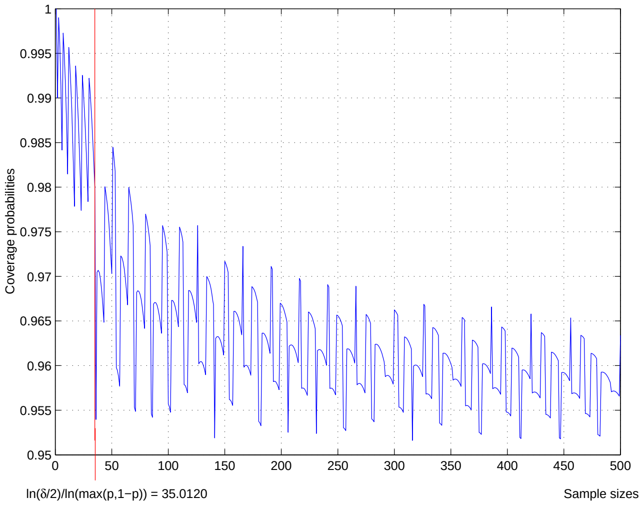

Fig. 1. The 95% CPCI Coverage probabilities for proportion $p = 0.1$

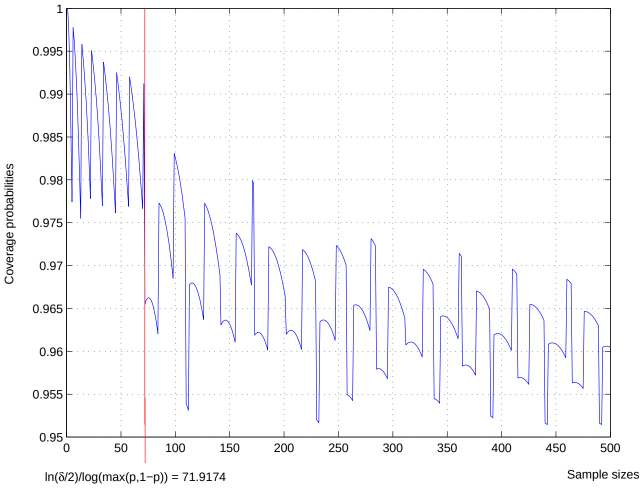

Fig. 2. The 95% CPCI Coverage probabilities for proportion $p = 0.05$

and

$$F_{u_k}(k) = \alpha/2 \text{ for every } k \in \{0, \ldots, n-1\}. \qquad (6)$$

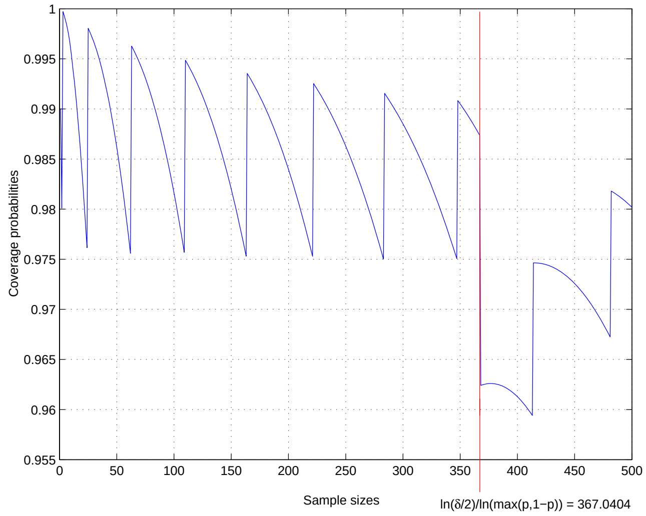

Fig. 3. The 95% CPCI Coverage probabilities for proportion $p = 0.01$

According to [1, Eq. 6.6.4, p. 263], for every given $\theta \in [0, 1]$,

$$G_\theta(k) = I_\theta(k, n - k + 1) \text{ for } k \in \{1, \dots, n\}, \quad (7)$$

and

$$F_\theta(k) = 1 - I_\theta(k + 1, n - k) \text{ for } k \in \{0, \dots, n - 1\}. \quad (8)$$

where $I_x(a, b)$ stands for the Incomplete Beta Function ([1, Eq. 6.2.1, p. 258, Eq. 6.6.2, p. 263, Eq. 26.5.1, p. 944]). The maps $\theta \in [0, 1] \longmapsto I_\theta(k + 1, n - k)$ and $\theta \in [0, 1] \longmapsto I_\theta(k, n - k + 1)$ are strictly increasing. Thereby, we can straightforwardly state the following result.

Lemma 3.1: Given $\theta \in [0, 1]$,

(i) *for every* $k \in \{0, \dots, n-1\}$, *the map* $\theta \in [0, 1] \longmapsto F_\theta(k) \in [0, 1]$ *is strictly decreasing;*

(ii) *for every* $k \in \{1, \dots, n\}$, *the map* $\theta \in [0, 1] \longmapsto G_\theta(k) \in [0, 1]$ *is strictly increasing.*

The subsequent lemma states useful properties concerning the real values ℓ_k and u_k.

Lemma 3.2: With the same notations as above,

(i) *the sequence* $(\ell_k)_{k\in\{0,\dots,n\}}$ *is strictly increasing and* $\ell_n = (\alpha/2)^{1/n}$;

(ii) *the sequence* $(u_k)_{k\in\{0,\dots,n\}}$ *is strictly increasing and* $u_0 = 1 - (\alpha/2)^{1/n}$.

(iii) *for every given* $k \in \{0, \dots, n\}$, $\ell_k < u_k$.

PROOF:

Proof of statement (i). Given $k \in \{1, \dots, n\}$, $\ell_k \in (0, 1)$ and is therefore larger than $\ell_0 = 0$. It then suffices to prove that $\ell_k < \ell_{k+1}$ for $k \in \{1, \dots, n - 1\}$ to establish the strict increasingness of the sequence $(\ell_k)_{k\in\{0,\dots,n\}}$.

Let $k \in \{1, \dots, n - 1\}$. We have that $G_{\ell_{k+1}}(k + 1) < G_{\ell_{k+1}}(k)$. According to (5), $G_{\ell_{k+1}}(k + 1) = G_{\ell_k}(k)$. It thus follows that $G_{\ell_k}(k) < G_{\ell_{k+1}}(k)$. According to lemma 1, statement (ii), we can conclude that $\ell_k < \ell_{k+1}$.

For every $\theta \in [0, 1]$, $G_\theta(n) = \theta^n$. It then follows from (5) that $\ell_n^n = \alpha/2$, which completes the proof of statement (i).

Proof of statement (ii). The strict increasingness of the sequence $(u_k)_{k\in\{0,\dots,n\}}$ derives from the same type of arguments as those used for proving statement (i). Given $k \in \{0, \dots, n - 1\}$, u_k belongs to $(0, 1)$ and is therefore less than $u_n = 1$. It then suffices to show that $u_k < u_{k+1}$ for every $k \in \{0, \dots, n - 2\}$ to establish the strict increasingness of the sequence.

Let $k \in \{0, \dots, n - 2\}$. We have that $F_{u_{k+1}}(k + 1) > F_{u_{k+1}}(k)$. According to (6), $F_{u_{k+1}}(k + 1) = F_{u_k}(k)$. Therefore, $F_{u_k}(k) > F_{u_{k+1}}(k)$. The result then follows from lemma 1, statement (i).

Given $\theta \in (0, 1)$, $F_\theta(0) = (1 - \theta)^n$. Therefore, by (6), we obtain that $F_{u_0}(0) = (1 - u_0)^n = \alpha/2$.

Proof of statement (iii). According to (1), (3) and the values of ℓ_n and u_0, statement (iii) holds true for $k = 0$ and $k = n$.

Consider any $k \in \{1, \dots, n-1\}$. We can write that $F_{\ell_k}(k-1) = 1 - G_{\ell_k}(k) = 1 - (\alpha/2) > \alpha/2$. This follows from the definition of G_θ, (5) and the fact that $\alpha < 1$. Since $F_{\ell_k}(k) > F_{\ell_k}(k - 1)$ and $\alpha/2 = F_{u_k}(k)$ according to (6), we finally obtain that $F_{\ell_k}(k) > F_{u_k}(k)$. The fact that $\ell_k < u_k$ then straightforwardly derives from lemma 1, statement (i).

■

Lemma 3.3: With the same notations as above,

(i) *if* $p = 0$, $P\left(\{\ell_X \geq p\}\right) = 1$,

(ii) *if* $0 < p \leq (\alpha/2)^{1/n}$, $P\left(\{\ell_X \geq p\}\right) \leq \alpha/2$

(iii) *if* $(\alpha/2)^{1/n} < p \leq 1$, $P\left(\{\ell_X \geq p\}\right) = 0$.

PROOF: Statement (i) is straightforward since $\ell_X \geq 0$.

As far as statements (ii) and (iii) are concerned, it is convenient to introduce the set $\mathcal{E} = \{k \in \{0, \dots, n\} : \ell_k \geq p\}$. Statement (i) of lemma 2 implies the following facts. First, $\mathcal{E}$ is non empty if and only if $p \leq (\alpha/2)^{1/n}$; second, if $0 < p \leq (\alpha/2)^{1/n}$, $\mathcal{E} = \{m, \dots, n\}$ where $m \geq 1$ according to (1).

Thereby, $\ell_X \geq p$ if and only if $X \geq m$. We therefore have that $P(\{\ell_X \geq p\}) = G_p(m)$. Now, since $p \leq \ell_m$, it follows from lemma 1, statement (ii), that $G_p(m) \leq G_{\ell_m}(m)$. According to (5), the right hand side in this inequality equals $\alpha/2$. Therefore, we obtain that $P(\{\ell_X \geq p\}) \leq \alpha/2$.

If $(\alpha/2)^{1/n} < p \leq 1$, $\mathcal{E}$ is empty. Therefore, $\ell_X < p$ and statement (iii) follows.

■

Lemma 3.4: With the same notations as above,

(i) *if* $0 \leq p < 1 - (\alpha/2)^{1/n}$, $P\left(\{u_X \leq p\}\right) = 0$,

(ii) *if* $1 - (\alpha/2)^{1/n} \leq p < 1$, $P\left(\{u_X \leq p\}\right) \leq \alpha/2$,

(iii) *if* $p = 1$, $P\left(\{u_X \leq p\}\right) = 1$.

PROOF: Set $\mathcal{F} = \{k \in \{0, \dots, n\} : u_k \leq p\}$. It follows from statement (ii) of lemma 2 that $\mathcal{F}$ is empty if and only if $0 \leq p < 1 - (\alpha/2)^{1/n}$. Therefore, if $0 \leq p < 1 - (\alpha/2)^{1/n}$, u_X is larger than p and statement (i) holds true.

Under the condition $1 - (\alpha/2)^{1/n} \leq p < 1$, $\mathcal{F}$ is not empty. According to statement (ii) of lemma 2 and (3), we obtain that $\mathcal{F} = \{0, \dots, L\}$ with $L \leq n - 1$. Therefore, the event $\{u_X \leq p\}$ is the event $\{X \leq L\}$ so that $P(\{u_X \leq p\}) = F_p(L)$. Since $u_L \leq p$, it follows from statement (i) of lemma 1 that

$F_p(L) \leq F_{u_L}(L)$. According to (6), the right hand side in this inequality equals $\alpha/2$ and statement (ii) follows.

Statement (iii) holds true since $u_X \leq 1$.

■

We now complete the proof of proposition 2.1. According to lemma 3.2, statement (iii), $\ell_X < u_X$. Thereby, $\{p \leq \ell_X\} \subset \{p < u_X\}$ so that

$$\begin{array}{rcl} P(\{\ell_X < p < u_X\}) & = & P(\{p < u_X\}) - P(\{p \leq \ell_X\}) \\ & = & 1 - P(\{p \geq u_X\}) - P(\{p \leq \ell_X\}) \end{array}$$

Since the condition $1 - (\alpha/2)^{1/n} \leq (\alpha/2)^{1/n}$ is equivalent to $n \geq -\ln(\alpha/2)/\ln(2)$, proposition 1 straightforwardly follows from lemmas 3.3 and 3.4.

IV. Conclusion

Because of the discreteness of the binomial distribution, the coverage probability of the $100(1-\alpha)\%$ CPCI for estimating a binomial parameter p is larger than or equal to $1-\alpha$. In addition to this standard result, this paper proves that this coverage probability is, in fact, larger than or equal to $1-\alpha/2$ when the sample size is less than a specific bound. Because this bound depends on the proportion to estimate, the coverage probability of the CPCI can be much larger than the confidence level. This is a major drawback of the CPCI. Therefore, the CPCI is not suitable for applications where it is important to form a confidence interval whose coverage probability is close to the specified confidence level.

References

[1] Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, Ninth printing.

[2] Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions, *The American Statistican*, **52**, NO. 2, 119-126.

[3] Brown, L. D. and Cai, T. and DasGupta, A. (2001). Interval Estimation for a Binomial Distribution, *Statist. Sci.*, **16**, 101-133.

[4] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence of fiducial limits illustrated in the case of the binomial, *Biometrika*, **26**, 404-413.

[5] Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics, Discrete Distributions*, John Wiley & Sons.

[6] Stuart, A. and Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics*, Fifth Edition, Charles Griffin.