Boletín de Estadística e Investigación Operativa
Vol. 27, No. 1, Febrero 2011, pp. 6-13

Estadística

A note about the chaotic behavior of the Wald Interval for a binomial proportion

Francisco J. Ortega, Jesús Basulto and José A. Camúñez

Departamento de Economía Aplicada I
Universidad de Sevilla
✉ fjortega@us.es, basulto@us.es, camunez@us.es

Abstract

It is well known that the coverage probability of the standard Wald confidence interval to estimate a binomial proportion has a very erratic behavior as a function of the parameters n (sample size) and p (probability of success) . Till now it has been thought that this behavior was "basically unpredictable". Nevertheless, the analysis of this behavior allows to obtain a formula that provides, for a fixed p, all the sample sizes in which the coverage probability decreases sharply from *n-1* to n.

Keywords: Binomial distribution, Confidence interval, Coverage probability.
AMS Subject classifications: 62F25.

1. Introduction

To obtain a confidence interval for the probability of success in a binomial distribution, one of the choices more widely used is the standard confidence interval based on normal approximation, usually so called Wald interval. Let us consider a simple random sample $X_1, ..., X_n$ from the Bernoulli distribution with parameter p (where p is the probability of success) and $X = \sum_{i=1}^{n} X_i$. It is well known that X is a binomial random variable with parameters n and p. The interval, of course, is $\widehat{p} \pm z_\alpha n^{-1/2}(\widehat{p}(1-\widehat{p}))^{1/2}$, where $\widehat{p} = X/n$ is the sample proportion of successes and z_α is the $100(1-\alpha/2)$th percentile of the standard normal distribution. The nominal confidence level of this interval is $1-\alpha$. This definition is easy to present, and is usually justified on the basis of the central limit theorem. In addition, it can be obtained from the Wald large-sample normal test. So at first glance, we may think that the problem is simple and the Wald interval is a solution totally satisfactory.

Nevertheless, the problem is really complex, because of the discrete nature of the binomial distribution. It has been pointed out that the coverage probability of the interval is often very far from the nominal confidence level. In fact, the majority of textbooks warn that this interval should be used only when certain conditions are fulfilled. Although the qualifications with which the standard interval is presented are varied, perhaps the most common is $n \cdot min\{p, 1-p\} \geq 5$ (or 10). This kind of condition is concerned about the poor coverage of the interval when p is near the boundaries 0 or 1.

Really, the problem of the Wald interval's coverage probability is far deeper. In Brown et al. (2001) there are several references to articles in which it has been pointed out that the coverage properties of the standard interval can be erratically poor even if p is not near the boundaries and the authors conclude that this behavior is more persistent than the statisticians have appreciated till now. In addition, the problem does not get away even when n is quite large.

In this article, we will focus on analyzing the behavior of the coverage probability as function of n, when $(p \leq 0.5)$ is fixed. For this situation, Brown et al. (2001) shows that there exist some "lucky" pairs (n, p) such that the actual coverage probability is very close to the nominal level, and other "unlucky" pairs (n, p) such that the corresponding coverage is much smaller than the nominal level. For instance, when $p = 0.05$, the actual coverage probability of the nominal 95% interval is 0.953 if $n = 17$, but falls to 0.919 when $n = 40$. When p is near to 0, this erratic behavior is more persistent and disconcerting. For instance, when $p = 0.005$ (and the nominal confidence is 95%), the coverage probability increases monotonically in n to the level 0.945 when $n = 591$ and then decreases dramatically to 0.792 if $n = 592$. The same behavior happens from $n = 953$ to $n = 954$, from $n = 1278$ to $n = 1279$, and on and on.

At first glance, the unlucky n appears in an unpredictable way. For instance, in Brown et al. (2001), p.102, we can read:

> "...the coverage of the standard interval can be significantly lower at quite large samples sizes, and this happens in an unpredictable and rather random way."

The main objective of our paper is to analyze, for fixed $p \leq 0.5$, why these sharp decreases happen in the coverage probability and to provide a formula to obtain all the "unlucky" values of n for which this occurs, without the need to calculate directly the coverage probability for all n.

Specifically, we have found an "empirical rule" from which we can deduce the formula that allows us to obtain the values of n in which the coverage probability decreases when we rise the sample size from *n-1* to *n*. We have found that this rule is verified, without exception, considering a wide range of values of n and p, although we could not have demonstrated it formally.

In Section 2, we present the standard Wald interval and the formula that allows us to calculate the actual coverage probability of this interval. In Section 3, we analyze the behavior of the coverage probability through some examples and we establish the empirical rule. In Section 4, we present the formula that allows us to obtain the "unlucky" values of n. Finally, in Section 5 we indicate some concluding remarks.

2. The Wald Interval and its coverage probability

Let us consider X a binomial random variable with parameters n and p (where n is the sample size and p is the probability of success). We want to obtain a confidence interval (CI) for the unknown parameter p with a confidence level $1-\alpha$, where α is some specified value between 0 and 1. Because of the discrete nature of the binomial model, we know that it is not possible to obtain a nonrandomized confidence interval that always achieves the exact nominal confidence level. If we want to consider only nonrandomized intervals, the most that we can achieve is that the coverage probability is "approximately" $1-\alpha$, that is, $P_p[p \in CI] \simeq 1-\alpha$. Following Brown et al. (2001) we will use the notation $C(n,p) = P_p[p \in CI]$ for the coverage probability.

One of the most widely used choices is the so called Wald interval. As we noted in Section 1, the interval is

$$\widehat{p} \pm z_\alpha n^{-1/2}(\widehat{p}(1-\widehat{p}))^{1/2}, \tag{2.1}$$

where $\widehat{p} = X/n$ is the sample proportion of successes and $z_\alpha = \Phi^{-1}(1-\alpha/2)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. This interval is obtained from the pivotal quantity

$$\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}},$$

whose asymptotic distribution is a standard normal distribution.

Thus, we can guarantee that for any fixed $p \in (0,1)$, $\lim_{n\to\infty} C(n,p) = 1-\alpha$. However, it is very important to emphasize that fixed p, the coverage probability is not monotonically increasing in n, that it is very far from the nominal level for some values of n and that the problem does not go away even when n is quite large.

From the definition of the interval, we can calculate its coverage probability by straightforward calculation. Specifically, the probability is

$$C(n,p) = \sum_{L_1(n,p)\leq j \leq L_2(n,p)} P[X=j], \tag{2.2}$$

where $L_1(n,p)$ and $L_2(n,p)$ are the solutions (in l) of the equations $n^{-1}(l + z_\alpha(l(n-l)n^{-1})^{-1/2} = p$ and $n^{-1}(l - z_\alpha(l(n-l)n^{-1})^{-1/2} = p$, respectively.

We can easily solve these equations and we obtain that

$$L_1(n,p) = \frac{n(z_\alpha^2 + 2np) - z_\alpha n\sqrt{z_\alpha + 4np(1-p)}}{2(z_\alpha^2 + n)}, \tag{2.3}$$

$$L_2(n,p) = \frac{n(z_\alpha^2 + 2np) + z_\alpha n\sqrt{z_\alpha + 4np(1-p)}}{2(z_\alpha^2 + n)}, \tag{2.4}$$

as we can see in Brown et al. (2002). For instance, when $p = 0.5$ and $n = 17$, we obtain $L_1(17, 0.5) = 4.8508$ and $L_2(17, 0.5) = 12.1492$, thus the coverage probability is given by $\sum_{j=5}^{12} P[X = j] = 0.9510$, taking into account that X is a binomial random variable with parameters $n = 17$ and $p = 0.5$.

3. The reason for the chaotic behavior of the coverage probability

As we have pointed out in Section 1, the coverage probability of standard Wald confidence interval has a very erratic behavior as a function of the parameters n and p. We will focus our attention on the analysis of such behavior when n increases and $p \leq 0.5$ is fixed.

Example 1: Figure 1 shows the coverage probability of the nominal 95% interval for fixed $p = 0.25$ and variable n from 20 to 60. We can appreciate that there are many “unlucky” values of n, in which the coverage probability falls sharply, and such values arise suddenly. In our example, these values are $\{25, 31, 37, 42, 48, 53, 58\}$. Let us emphasize that, from now, we will say specifically that a value of n is “unlucky” if $C(n,p) < C(n-1,p)$. Though our interest focusses on the sharp decreases, it is interesting to remark that there is a systematic negative bias in the coverage probability, since it is almost always less than the nominal level $1 - \alpha = 0.95$.

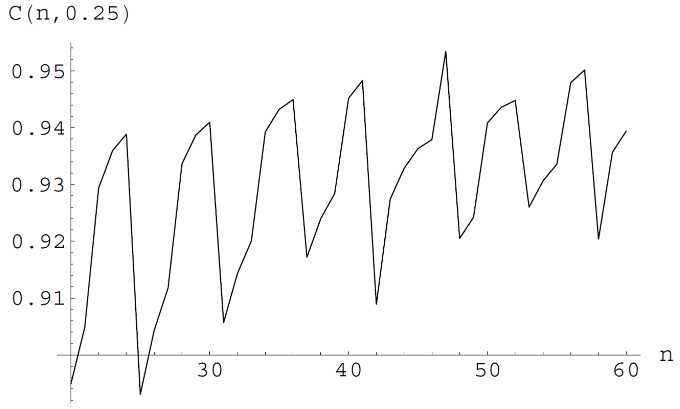

Figure 1: Coverage probability for fixed $p = 0.25$ and variable $n = 20$ to 60

From the formula (2.2) that allows to calculate $C(n,p)$, it is easy to understand at an intuitive level that the oscillation in the coverage probability is caused by the discreteness of the binomial model. Indeed, given the values of n and p the coverage probability of the interval is the sum of some of all possible values of a binomial random variable with parameters n and p. Concretely, we must add only the probabilities of the integer values in the interval $[L_1(n,p), L_2(n,p)]$, that is, the first term is the smallest integer larger than or equal to $L_1(n,p)$, namely $\ell_{n,p}$, and the last term is the largest integer smaller than or equal to $L_2(n,p)$, namely $u_{n,p}$. Thus, as we can see in Brown et al. (2002), p.167, what happens is that a small change in n or p can cause $\ell_{n,p}$ and/or $u_{n,p}$ to leap to the next integer value.

Let us consider, for instance, the case $p = 0.25$ y $\alpha = 0.05$. For $n = 41$, we have $L_1(41, 0.25) = 5.858$ and $L_2(41, 0.25) = 16.398$ and hence $\ell_{41,0.25} = 6$ and $u_{41,0.25} = 16$; when $n = 42$, we have $L_1(42, 0.25) = 6.042$ and $L_2(42, 0.25) = 16.718$ and therefore $\ell_{41,0.25}$ increases to 7, while $u_{42,0.25}$ remains 16. Thus, when n increases, the sum loses a term and this fact implies the decrease of the coverage probability.

4. The formula to obtain the "unlucky" values

Let us consider again the case $p = 0.25$. Table 1 list the values of $L_1(n, 0.25)$, $L_2(n, 0.25)$, $\ell_{n,0.25}$, $u_{n,0.25}$ and $C(n, 0.25)$ for some values of n. We have highlighted with dark background the "unlucky" values of n . Let us remark that both L_1 and L_2 are strictly increasing in n (it is easy to demonstrate this property calculating the derivatives of both functions and checking that they are positives). On the other hand, we can verify that *the probability decreases when, and only when, the integer part of $L_1(n, 0.25)$ increases.*

We have seen empirically for a wide range of values of n and p that the above property is always verified, without any exception. Thus, we will establish the following "empirical rule": *Fixed $p \leq 0.5$, when we rise from $n-1$ to n, the coverage probability of the Wald interval decreases if and only if the integer part of L_1 increases*, that is,

$$C(n,p) < C(n-1,p) \Leftrightarrow \ell_{n,p} > \ell_{n-1,p}. \tag{4.1}$$

This empirical rule (or conjecture) allows to obtain all the values of n in which the coverage probability decreases when we rise from $n-1$ to n. Indeed, solving (in n) the equations

$$L_1(n,p) = k, \ \ k \in N, \tag{4.2}$$

let us consider $\{n_k^*\}_{k\in N}$ the set of solutions of such equations. Then, the set of "unlucky" values of n is given by $\{n_k = Int[n_k^*] + 1\}_{k\in N}$, where $Int[z]$ is the

Table 1: Extremes terms of the sum and coverage probability for fixed $p = 0.25$.

n	$L_1(n, 0.25)$	$L_2(n, 0.25)$	$\ell_{n,0.25}$	$u_{n,0.25}$	C(n,0.25)
23	2.7164	10.4294	3	10	0.9359
24	2.8799	10.7759	3	10	0.9389
25	3.0450	11.1200	4	11	0.8931
26	3.2116	11.4618	4	11	0.9043
...	...	...	...	...	...
30	3.8926	12.8101	4	12	0.9410
31	4.0661	13.1428	5	13	0.9057
...	...	...	...	...	...
36	4.9502	14.7854	5	14	0.9449
37	5.1300	15.1101	6	15	0.9172
...	...	...	...	...	...
41	5.8581	16.3980	6	16	0.9483
42	6.0422	16.7176	7	16	0.9089

integer part of z. One option to obtain these values of n is to solve numerically the equation (4.2) using an appropriate software. However, if we square the equation (4.2), we obtain a fourth degree equation; with help of the software Mathematica 5.2, we can confirm that this equation has only one real and positive solution that verifies the initial equation (4.2). Specifically, defining

$$\begin{aligned} f(p, z_\alpha, k) &= k^2p^2 + 3kp^2z_\alpha^2 && (4.3)\\ g(p, z_\alpha, k) &= -2k^3p^3 + 18k^2p^3z_\alpha^2 - 27k^2p^4z_\alpha^2 && (4.4)\\ h(p, z_\alpha, k) &= \sqrt{4f(p, z_\alpha, k)^3 - g(p, z_\alpha, k)^2}, && (4.5) \end{aligned}$$

we obtain that the solution of equation (4.2) in which we are interested is given by

$$n_k^* = \frac{2k}{3p} + \frac{2\sqrt{f(p, z_\alpha, k)}}{3p^2} \cos\left(\frac{1}{3}\arctan\left(\frac{h(p, z_\alpha, k)}{g(p, z_\alpha, k)}\right)\right), \qquad (4.6)$$

and therefore, the succession of "unlucky" values of n, fixed p, is

$$\left\{n_k = Int\left[\frac{2k}{3p} + \frac{2\sqrt{f(p, z_\alpha, k)}}{3p^2} \cos\left(\frac{1}{3}\arctan\left(\frac{h(p, z_\alpha, k)}{g(p, z_\alpha, k)}\right)\right)\right] + 1\right\}_{k\in N}. \qquad (4.7)$$

Let us remark that applying the general formula to solve fourth degree equations, the initial solution given by Mathematica 5.2 is

$$n_k^* = \frac{2k}{3p} + \frac{2^{1/3}f(p, z_\alpha, k)}{3p^2 i(p, z_\alpha, k)} + \frac{i(p, z_\alpha, k)}{3p^2 2^{1/3}},$$

Table 2: Extremes terms of the sum and coverage probability for fixed $p = 0.005$.

n	$L_1(n, 0.005)$	$L_2(n, 0.005)$	$\ell_{n,0.005}$	$u_{n,0.005}$	C(n,0.005)
2155	5.9863	19.3600	6	19	0.9508
2156	5.9898	19.3665	6	19	0.9502
2157	5.9934	19.3729	6	19	0.9503
2158	5.9970	19.3793	6	19	0.9504
2159	6.0006	19.3857	7	19	0.9056
2160	6.0041	19.3922	7	19	0.9057
2161	6.0077	19.3986	7	19	0.9059

where

$$i(p, z_\alpha, k) = \left(g(p, z_\alpha, k) + \sqrt{-4f(p, z_\alpha, k)^3 + g(p, z_\alpha, k)^2}\right)^{1/3}.$$

In this expression, $i(p, z_\alpha, k)$ is a complex number, because of $-4f(p, z_\alpha, k)^3 + g(p, z_\alpha, k)^2$ is less than 0. But we can see that the second and third fractions in this formula of n_k^* are conjugate complex numbers. Thus, an alternative formula is

$$n_k^* = \frac{2k}{3p} + 2Re\left[\frac{i(p, z_\alpha, k)}{3p^2 2^{1/3}}\right],$$

where $Re[z]$ is the real part of the complex number z. Calculating $Re\left[i(p, z_\alpha, k)\right]$ and replacing it in the last formula, we obtain finally (4.6).

Consider $p = 0.25$. The first values given by formula (4.7) are {12, 19, 25, 31, 37, 42, 48, 53, 58, 63, 68, ...}. In Example 1, with variable n from 20 to 60, we obtained the "unlucky" values {25, 31, 37, 42, 48, 53, 58}. We can verify the consistency between the two series; moreover, without calculating the coverage probabilities we can state that the next "unlucky" values are 63 and 68. Brown et al. (2001) consider the case $p = 0.005$, and they obtain calculating the coverage probability the first five "unlucky" values {592, 954, 1279, 1583, 1876}. The first seven terms given by (4.7) in this case are {592, 954, 1279, 1583, 1876, 2159, 2436}, and then we can deduce that the next "unlucky" value is $n = 2159$. In fact, Table 2 list the values of $L_1(n, 0.005)$, $L_2(n, 0.005)$, $\ell_{n,0.005}$, $u_{n,0.005}$ and $C(n, 0.005)$ in the neighbor of $n = 2159$, and we can confirm that the probability fall occurs in this value of n due to the rise in the integer part of L_1.

5. Concluding remarks

Interval estimation of the probability of success in a binomial distribution is a very basic but very important problem of statistical practice. As it has been pointed out in Brown et al. (2001,2002), and references therein, the coverage probability of the most widely used interval, namely the standard Wald interval, have a chaotic behavior. Besides, in many cases, this erratic behavior does not go away even when n is quite large. The authors recommend other alternative intervals.

The sharp oscillations in the coverage probability is caused by the discreteness of the binomial model and, at first glance, happens in an unpredictable and random way. We observed a property of the coverage probability of the Wald interval that helps us to understand why the coverage probability decreases sometimes when we increase from $n-1$ to n (for fixed $p \leq 0.5$), and from this property we can deduce a formula to obtain all these "unlucky" values of n, without need to calculate the coverage probabilities.

Let us remark that the restriction $p \leq 0.5$ does not reduce generality to the result, since we can always define the success of the experiment so that its probability p is smaller than or equal to 0.5.

In future researches, the most interesting thing would be to obtain a formal proof of the empirical rule given in (4.1).

References

[1] Brown L.D., Cai T., and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion), *Stat Sci*, **16**, 101-133.

[2] Brown L.D., Cai T., and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions, *Ann Stat*, **30**, 160-201.

About the authors

Fco. Javier Ortega is bachelor in Mathematics, doctor in Economics and professor of Statistics and Econometrics at the University of Sevilla. His main research lines are Bayesian Inference methods (in particular nonregular models) and History of the Probability and Statistics.

Jesús Basulto is full professor of Statistics at the University of Sevilla. His main research line is History of the Probability and Statistics.

José Antonio Camúñez is bachelor in Mathematics, doctor in Economics and professor of Statistics and Econometrics at the University of Sevilla. His main research lines are Multivariate Data Analysis and History of the Probability and Statistics.