Confidence Interval forBinomial Proportion

Sumana AL-Saadi

AL-Mustanisirya University

Abstract

The accuracy of a confidence interval expression is one of most important problems in statistics. In this paper we explore ways of computing a confidence interval for a binomial parameter. The most common formula, the normal approximation interval doesn't work well when the value of the proportion is small and we show that in the example. We determine if there is a method to produce a confidence interval where the true coverage probability is as close as possible to the target. In addition, we provide examples of how these methods are used in science and social applications.

Keywords: confidence interval; binomial distribution; binomial proportion; normal approximation interval; Wald confidence interval; coverage probability.

Introduction

The confidence interval estimation for binomial proportion is a subject debated in many scientific articles. A confidence interval for population parameter consists of a range of values restricted by a lower and an upper limit. The size of the interval depends on the sample size and on the confidence coefficient $1\Pi/7.$ **[3] Boomsma** (2005)

The normal approximation interval is one way that is used to compute the confidence interval for binomial proportion. Its simple formula is based on approximating the binomial distribution with a normal distribution by the central limit theorem. The formula is:

$$p^{\wedge} \pm z_{\alpha/2}\sqrt{\frac{p^{\wedge}(1-p^{\wedge})}{n}} \quad \ldots\ldots(1)$$

Where $p^{\wedge}$ **is the proportion of successes is estimated from the sample,** $z_{\alpha/2}$ **is the** $\alpha/2$ **percentile of a standard normal distribution and n is the sample size.**

This way fails totally when the sample proportion is exactly zero or one and the population proportion is in $(0,\ 1)$**. There are several competing formulae that work well especially for situations with a small sample size and a proportion very close to zero or one.**

Objective

There is no exact confidence interval for the binomial parameter. The aim of this research is determine if there is a method to produce confidence interval where true coverage probability is as close as possible to the target and explore the ways of computing a

confidence interval for a binomial parameter. In addition, provide examples of how these methods are used in science and social applications.

Literature Review

This paper describes the history of the development of the confidence interval that is most used and relied upon by statisticians now.

The normal approximation interval was first developed by [9] Wilson (1927). This interval has good properties even for a small number of trials and/or an extreme probability.

[2] Bohning (1994) discussed five methods described as Method I, II, III, IV and V for constructing approximate confidence intervals for the binomial parameter p. The results of comparison of these methods in respect to coverage probability and expected length appear to indicate that Method V has an advantage over Chen's Bayes method as well as over the other three methods.

[1] Agresti and Coull (1998) expanded the way used by statisticians to compute the confidence interval for a binomial proportion by using an asymptotic normality of the sample proportion and estimating the standard error.

Agresti and Coull called this the Wald confidence interval for p. Also, they described a way used to prevent approximation that was suggested by Clopper–Pearson's (1934) "exact" confidence interval for p. This way has coverage probabilities bounded below by the nominal confidence level, but the typical coverage probability is much higher than that level.

[7] Henderson and Meyer (2001) explained the problem faced by statisticians when they have a binomial distribution and need to estimate a binomial proportion using two ways: first, frequentist and second, Bayes methods. They also described how to compute confidence interval and compared other methods with Wilson interval (1927), which was explained earlier in this paper.

[4] Brown, Cai and Das Gupta, (2001) explained that other literature remarked on the erratic behavior of the coverage probability of the standard Wald confidence interval. So, this problem led the authors to consider alternative intervals. Also, they examined each interval for its coverage probability and its length based on this analysis. In case of a small n, they recommended the Wilson interval or the equal tailed Jeffreys prior interval but for a large n, they suggested the Agresti and Coull interval.

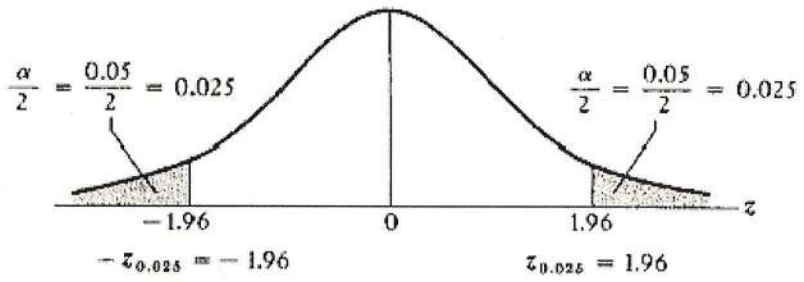
[8] Song, Chang and Liu (2009) explained that the Wilson procedure was thought to be superior to many existing procedures because it was less sensitive to p and less costly because reduce the sample size. They gave some procedures that work as well as the

Wilson procedure in the case that p is close to 0 or 1. Also, these are less sensitive than Wilson procedure. Song et al. gave the same coverage probabilities that Wilson gave when he used the sample size of 1021, but they only needed a sample size of 177.

Methodology

I-Definition of Confidence Interval

There are two types of estimates for each population parameter: the point estimate and confidence interval estimate. A point estimate is a single value given as the estimate of a population parameter that is of interest, for example the mean of some quantity. An interval estimate specifies instead a range within which the parameter is estimated to lie. Confidence intervals are commonly reported in tables or graphs along with point estimates of the same parameters, to show the reliability of the estimates.



The level of confidence of the confidence interval would indicate the probability that the confidence range captures this true population parameter given a distribution of samples. (6)

In applied practice, confidence intervals are typically stated at the 95% confidence level.(10) However, when presented graphically, confidence intervals can be shown at several confidence levels, for example 50%, 95% and 99%.

A confidence interval is an indicator of your measurement's precision. It is also an indicator of how stable your estimate is, which is the measure of how close your measurement will be to the original estimate if you repeat your experiment.

II-Confidence Interval Calculation

(5) a Bernoulli random variable Xi is defined to have two possible values: Success (Xi =1, with probability p), and Failure (Xi =0, with probability q = 1-p). A binomial random variable X is defined as the sum of n independent Bernoulli random variables (X1,...,Xn). The estimator for the population proportion is equal to X/n, the mean and variance of $\hat{p}$ are easily obtained:

$E[X/n] = (1/n)\, E[X] = np/n = p$ .....(2)

$V[X/n] = (1/n)^2\, V[X] = npq/n^2 = pq/n$.....(3)

Subtracting off the mean and standard deviation from $p^{\wedge}$ then gives a standard normal random variable .

The simplest and most commonly used formula for a binomial confidence interval (the normal approximation interval) is shown below:

$$CI = (p^{\wedge} - z_{\alpha/2}\partial^{\wedge}, p^{\wedge} + z_{\alpha/2}\partial^{\wedge}) \quad \ldots\ldots(4)$$

Where x is the number of successes. The estimate of p is

$$p^{\wedge} = x/n \quad \ldots\ldots(5)$$

The estimated standard error of the point estimator of p is

$$\partial^{\wedge} = \sqrt{\frac{p^{\wedge}(1-p^{\wedge})}{n}} \quad \ldots\ldots(6)$$

and $z_{\alpha/2}$ is the value satisfying $P(Z \le z_{\alpha/2}) = 1-(\alpha/2)$ where Z follows the standard normal distribution with mean 0 and variance 1.

As mentioned in the introduction, this way fails totally when the sample proportion is exactly zero or one. There are several competing formulae that work well especially for situations with a small sample size and a proportion very close to zero or one. One of them is Wilson confidence interval [9] Wilson (1927) that has the form:

$$CI_w = (p_w^{\wedge} - z_{\alpha/2}\partial_w^{\wedge}, p_w^{\wedge} + z_{\alpha/2}\partial_w^{\wedge}) \quad \ldots\ldots(7)$$

Where:

$$p_w^{\wedge} = \frac{x + (z_{\alpha/2}^2/2)}{n + z_{\alpha/2}^2} \quad \ldots\ldots(8)$$

and $$\partial_w^{\wedge} = \sqrt{\frac{np^{\wedge}(1-p^{\wedge}) + (z_{\alpha/2}^2/4)}{(n+z_{\alpha/2}^2)^2}} \quad \ldots\ldots(9)$$

where $p^{\wedge}$ was defined in Eq. (5).

Results

To apply this formula we need to use real data. So, we will use the data for two groups of patients a treatment group and a control group and the size of each one is equal to 302 and 303. The goal of this study is to test the effectiveness of a Mediterranean-type diet on the rate of coronary events in people who have had a first heart attack. After 27 months, the study found that there were 16 cardiac deaths from the control group and 3 cardiac deaths from the treatment group.

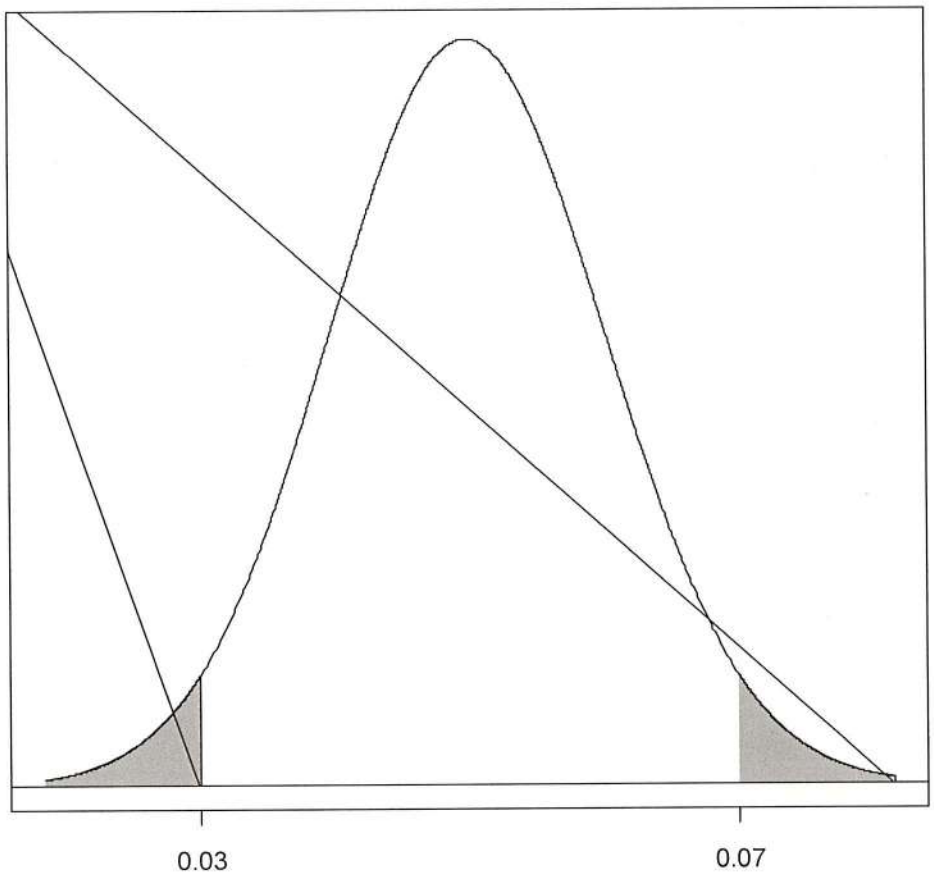
The resulting 95% (the normal approximation interval) would be calculated as follows:

$$\hat{p_t} = .05$$

$$C.I._t = .05 \pm 1.96(0.0125)$$

We are 95% confident that the true proportion of patients who have had a first heart attack is between 0.025 and 0.074 for the treatment group.

Figure 1.Coverage probabilities,p=.05.

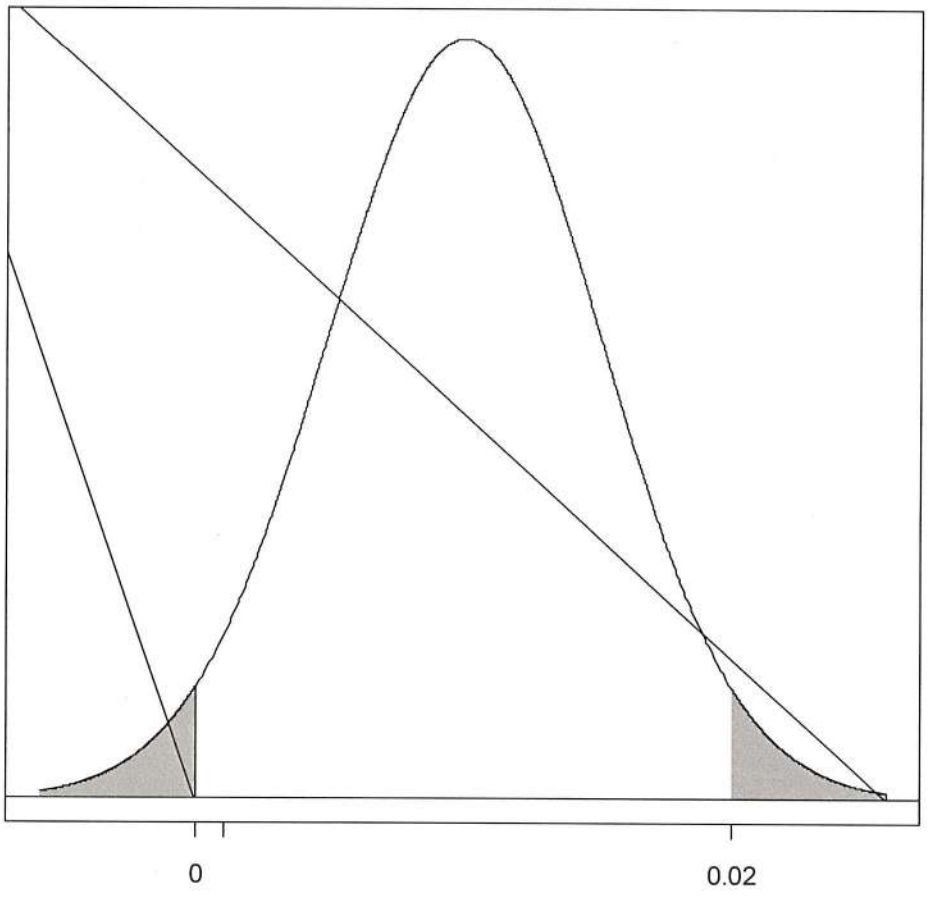


$$\hat{p_c} = .01$$

$$C.I._c = .01 \pm 1.96(0.0057)$$

We are 95% confident that the true proportion of patients who have had a first heart attack is between 0.001 and 0.021 for the control group.

Figure 2.Coverage probabilities,p=.01.



The normal approximation interval doesn't work well when the value of the proportion is small. So for this reason and others, scientists are trying to improve it. Another type of confidence interval that compares these values is described below:

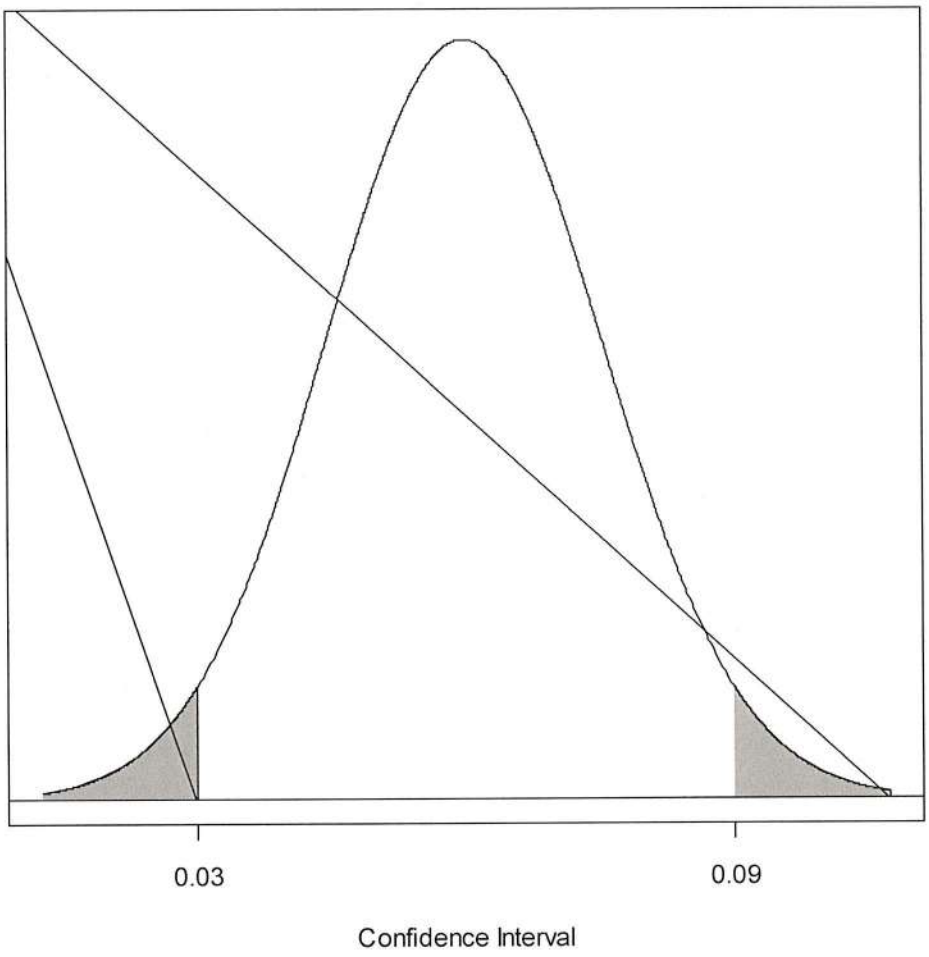
The resulting 95% (Wilson confidence interval) would be calculated as follows:

$$\hat{p_{w(t)}} = .06$$

$$C.I._{w(t)} = .06 \pm 1.96(0.0128)$$

We are 95% confident that the true proportion of patients who have had a first heart attack is between 0.034 and 0.085 for the treatment group.

Figure 3.Coverage probabilities,p=.06.

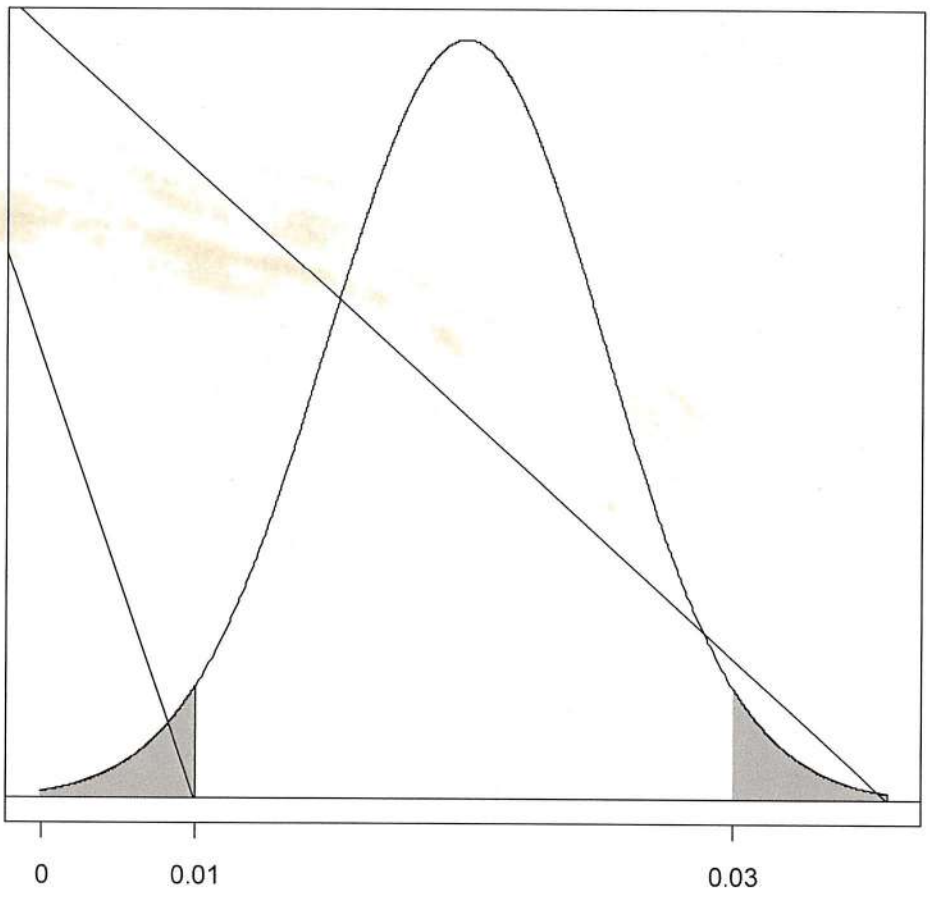


$$\hat{p}_{w(c)} = .02$$

$$C.I._{w(c)} = .02 \pm 1.96(0.0065)$$

We are 95% confident that the true proportion of patients who have had a first heart attack is between 0.007 and 0.032 for the control group.

Figure 4.Coverage probabilities,p=.02.



By using Wilson confidence interval smaller confidence interval was found. These findings improve the true value estimation.

References

1- Agresti , A., and Coull, B.A. (1998), *"Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions"* American Statistical Association, 52, 119-126.

2-Bohning, D. (1994),"*Better Approximate Confidence Intervals for a Binomial Parameter*" The Canadian journal of statistics, 22, 207-218.

3-Boomsma, A. (2005), "*Confidence Interval for Binomial Proportion*" Department of Statistics and Measurement Theory, University of Groningen.

4-Brown, L. D., Cai, T.T., and Das Gupta, A. (2001)," *Interval Estimation for a Binomial Proportion*" Statistical Science, 16, 101-117.

5-Dunnigan,K. (2008), *"Confidence Interval Calculation for Binomial Proportions"*. Statking Consulting, Inc.

6-Field, A. (2013),*"Discovering statistics using SPSS"*. London: SAGE

7-Henderson, M., and Meyer, M. (2001),*" Exploring the Confidence Interval for Binomial Parameter in a First Course in Statistical Computing"* The American statistician, 55, 337-344.

8-Song, W., Chang, C., and Liu, S. (2009*),"Robust Confidence Interval for the Bernoulli Parameter"* Communications in Statistics-Theory and Methods, 38, 3544-3560.

9-Wilson, E.B. (1927),*" Probable Inference, the Law of Succession, and Statistical Inference"* Journal of the American Statistical Association, 22, 209-212.

10-Zar, J.H. (1984), *"Biostatistical Analysis"*. Prentice Hall International, New Jersey. pp 43-45.

11- http://www.wikidata.org/wiki/Q208498#sitelinks-wikipedia

12- http://www.wikimediafoundation.org