



HAL
open science

An alternative method to compute confidence intervals for proportion

Antonio Francisco, Roldán López de Hierro, María del Mar López-Martín,
Carmen Batanero

► **To cite this version:**

Antonio Francisco, Roldán López de Hierro, María del Mar López-Martín, Carmen Batanero. An alternative method to compute confidence intervals for proportion. Eleventh Congress of the European Society for Research in Mathematics Education (CERME11), Utrecht University, Feb 2019, Utrecht, Netherlands. hal-02412839

HAL Id: hal-02412839

<https://hal.science/hal-02412839>

Submitted on 15 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An alternative method to compute confidence intervals for proportion

Antonio Francisco Roldán López de Hierro, María del Mar López-Martín and Carmen Batanero

University of Granada, Spain; aroldan@ugr.es, mariadelmarlopez@ugr.es, batanero@ugr.es

Keywords: Proportion, confidence level, confidence interval, interpretation.

Introduction

Confidence intervals are a main inferential procedure with many applications to real life, such as studying the percentage of voters will finally vote for a particular political party or the proportion of the population that suffers from asthma. This topic is taught in Spain to high school students in the Social Sciences specialty and in most of university grades. Since the available time for teaching is scarce, students are mainly taught the computing procedure, with no much attention to interpretation of results. Olivo, Batanero and Díaz (2008), and references therein, described some difficulties appearing in students' interpretation of confidence intervals for proportion.

The aim of this paper is to present an alternative methodology to compute confidence intervals for proportion by using the approach due to Wilson (1927). It would be of interest for students to discover that there is not a unique way to compute such interval. With the help of computers, by comparing Wilson's procedure to classical one, students would be able to improve their interpretation of the confidence interval for proportion and its confidence level.

Usual methodology to determine confidence intervals for proportions

Below, we summarize the usual procedure for computing the confidence interval. Suppose we are interested in estimating the proportion p of individuals in a population that satisfy certain condition. We take a random sample of the population, and observe the number X of individuals in the sample that verifies that property, which follows a Binomial distribution of parameters n and p . When n is large enough, we approximate this distribution by the normal distribution $X \approx \tilde{X} \mapsto N(np, npq)$, where $q = 1 - p$, and it can be typified as $(\hat{p} - p)/\sqrt{pq/n} \mapsto N(0,1)$ where \hat{p} is the sample proportion defined as the number of the successes divided by total number of observations in the sample. Let denote by $1 - \alpha$ the confidence level (usually $1 - \alpha \in \{0.9, 0.95, 0.99\}$) and let $z_{1-\alpha/2} > 0$ be the unique positive real number such that $P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$, where Z is any random variable following the standard normal distribution $N(0,1)$. Then the following approximation can be considered:

$$P\left(-z_{1-\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2}\right) = 1 - \alpha. \quad (1)$$

Since p is unknown, we replace it by its maximum likelihood estimator, \hat{p} , from which it is easy to deduce that the confidence interval for p , at the confidence level $1 - \alpha$, is

$$IC(p) = \left[\hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \right]. \quad (2)$$

The advantage of this expression is that it is very simple to use in practice and it produces good results. Its main weakness is that it has been obtained through a process in which, without apparent justification, the true value of p has been replaced by its estimator \hat{p} . This replacement transforms a

non-linear problem into a linear problem. One of the most important drawbacks of expression (2) is that this formula does not reach the confidence level.

Wilson’s methodology to determine the confidence interval for proportion

This method proposes to solve the equation (of second degree after squaring):

$$\frac{\hat{p} - p}{\sqrt{p(1 - p)/n}} = z_{1-\alpha/2}$$

before replacing p by \hat{p} in the denominator. Its two solutions determine the extremes of the confidence interval for proportion, which are:

$$IC_W(p) = \left[\frac{2n\hat{p} + z_{1-\alpha/2}^2 \pm z_{1-\alpha/2} \sqrt{z_{1-\alpha/2}^2 + 4n\hat{p}(1 - \hat{p})}}{2(n + z_{1-\alpha/2}^2)} \right]. \quad (3)$$

Since this formulae is more complex than (2), we propose to use a spreadsheet to generate a lot of random samples of distribution $B(n, p)$ (for instance, 1000 or more), when n and p are previously set by students. By using that random samples, we can compute both (2) and (3) intervals, and to determine how many of them indeed contain the true value of p . This comparison will help students to understand that the correct interpretation of confidence interval relies on the methodology, but not on the interval: when lots of intervals are randomly generated, we trust that $100(1 - \alpha)\%$ will contain the (unknown) true value of the proportion p . Furthermore, we are interested on overcoming the wrong student’s usual affirmation that establishes that, after computing the confidence interval, the probability that p belongs to such interval is the confidence level. This novel view-point could be of interest for both undergraduate and graduate students.

To compare both methods, we present the results of 10000 and 1000000 simulation with $n = 30$ (see Table 1). We highlight that only around the 88% of the intervals constructed using (2) contains the real value of p and the greater confidence of Wilson’s method.

Usual method		Wilson’s method	
$N = 10000$	$N = 1000000$	$N = 10000$	$N = 1000000$
87.19%	87.54%	94.88 %	95.23 %

Table 1: Proportion of confidence intervals in simulation containing the true value of p

Project EDU2016-74848-P (AEI, FEDER) and Research groups FQM126 y FQM268 (Junta de Andalucía).

References

- Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209–212.
- Olivo, E., Batanero, C. and Díaz, C. (2008). Dificultades de comprensión del intervalo de confianza en estudiantes universitarios. *Educación Matemática*, 20(3), 55–82.