Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On split sample and randomized confidence intervals for binomial proportions

Q1 Måns Thulin

Department of Mathematics, Uppsala University, Sweden

ARTICLE INFO

Article history:
Received 27 February 2014
Received in revised form 5 May 2014
Accepted 9 May 2014
Available online xxxx

Keywords:
Binomial distribution
Confidence interval
Randomization
Split sample method

ABSTRACT

We study randomized confidence intervals for binomial proportions, comparing coverage, length and the impact of the randomization. It is seen that the recently proposed split sample intervals can be improved upon in various ways. Criticism of randomized intervals are discussed.

1. Introduction

The problem of constructing confidence intervals for parameters of discrete distributions continues to attract considerable interest in the statistical community. The lack of smoothness of these distributions causes the coverage (i.e. the probability that the interval covers the true parameter value) of such intervals to fluctuate from $1 - \alpha$ when either the parameter value or the sample size n is altered. Recent contributions to the theory of these confidence intervals include Krishnamoorthy and Peng (2011), Newcombe (2011), Gonçalves et al. (2012), Göb and Lurz (2013) and Thulin (2013a).

The purpose of this short note is to discuss the split sample method recently proposed by Decrouez and Hall (2014). The split sample method reduces the oscillations of the coverage by splitting the sample in two. It is applicable to most existing confidence intervals for parameters of lattice distributions, including the binomial and Poisson distributions. For simplicity, we will in most of the remainder of the paper limit our discussion to the binomial setting, with the split sample method being applied to the celebrated Wilson (1927) interval for the proportion p. Our conclusions are however equally valid for other confidence intervals and distributions.

A random variable $X \sim \text{Bin}(n, p)$ is the sum of a sequence $X_1, \dots, X_n$ of independent Bernoulli(p) random variables. The split sample method is applied to $X_1, \dots, X_n$ rather than to X. The idea is to split the sample into two sequences $X_1, \dots, X_{n_1}$ and $X_{n_1+1}, \dots, X_{n_1+n_2}$ with $n_1 + n_2 = n$ and $n_1 \neq n_2$. In the formula for the chosen confidence interval, the maximum likelihood estimator $\hat{p} = X/n$ is then replaced by the weighted estimator

$$\tilde{p} = \frac{1}{2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i + \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} X_i\right). \tag{1}$$

The formula for the Wilson interval is

$$\frac{\hat{p}n + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} \pm \frac{z_{\alpha/2}}{n + z_{\alpha/2}^2}\sqrt{\hat{p}(1-\hat{p})n + z_{\alpha/2}^2/4},$$

E-mail address: thulin@math.uu.se.

http://dx.doi.org/10.1016/j.spl.2014.05.005
0167-7152/© 2014 Published by Elsevier B.V.

Please cite this article in press as: Thulin, M., On split sample and randomized confidence intervals for binomial proportions. Statistics and Probability Letters (2014), http://dx.doi.org/10.1016/j.spl.2014.05.005

where $z_{\alpha/2}$ is the $\alpha/2$ standard normal quantile, so the split sample Wilson interval is given by

$$\frac{\tilde{p}n + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} \pm \frac{z_{\alpha/2}}{n + z_{\alpha/2}^2}\sqrt{\tilde{p}(1-\tilde{p})n + z_{\alpha/2}^2/4}.$$

Decrouez and Hall (2014) showed by numerical and asymptotic arguments based on Decrouez and Hall (2013) that when $n_1 = n/2 + 0.15n^{3/4}$ (rounded to the nearest integer) and $n_2 = n - n_1$ the split sample method greatly reduces the oscillations of the coverage of the interval, without increasing the interval length. In the remainder of the paper, whenever n_1 and n_2 need to be specified, we will use these values.

Depending on how the sample is split, different confidence intervals will result, making the split sample interval a randomized confidence interval. If the sequence $X_1, \ldots, X_n$ is available, the interval can be *data-randomized* in the sense that the randomization can be determined by the data: the first n_1 observations can be put in the first subsample and the remaining n_2 observation in the second subsample. If the results of the individual trials have not been recorded, one must use randomness from outside the data to create a sequence $X_1, \ldots, X_n$ of 0's and 1's such that $\sum_{i=1}^n X_i = X$. We will refer to the latter strategy as *external randomization* and will discuss these two settings separately.

Decrouez and Hall (2014) left two questions open. The first is how the split sample interval performs in comparison to other randomized intervals, as Decrouez and Hall (2014) only compared the split sample interval to non-randomized confidence intervals. The second question is to what extent the randomization can affect the bounds of the interval. In the remainder of this note we answer these questions, discussing the impact of the random splitting on the confidence interval and comparing the split sample interval to alternative intervals. In Section 2 we describe a connection between split sample methods and adding discrete noise to the data. Section 3 is concerned with externally randomized intervals whereas we in Section 4 study data-randomized intervals. Various criticisms of randomized intervals are then discussed in Section 5.

2. Random splitting and discrete noise

A strategy for smoothing the distribution of the binomial random variable X is to base our inference on $X + Y$, where Y is a comparatively small random noise, using $X + Y$ instead of X in the formula for our chosen confidence interval. Having a smoother distribution leads to a better normal approximation, which in turn reduces the coverage fluctuations of the interval. From a purely probabilistic perspective, the split sample method can be seen to be a special case of this strategy. Let Z be a random variable which, conditioned on X, follows a Hypergeometric(n, X, n_1) distribution. Then it follows from (1) that

$$\tilde{X} = n\tilde{p} \overset{d}{=} \frac{n}{2n_1}Z - \frac{n}{2n_2}(X - Z),$$

so that

$$\tilde{X} \overset{d}{=} X + Y \quad \text{with } Y = \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n_1 - n_2}{2n_2}X.$$

The conditional distribution of Y when $n = 11$ is shown in Fig. 1.

From the above distributional identity it is clear that the split sample method relies on the sufficient statistic X as well as an additional random variable Y, and that it therefore can be considered to amount to adding discrete noise to binomial data. We note however that the noise term Y is a deterministic function of the sequence $X_1, \ldots, X_n$. Conditioned on the sequence, it is therefore more natural to think of $\tilde{X}$ as resulting from splitting the sample rather than adding noise to X.

3. Externally randomized confidence intervals

3.1. Other externally randomized intervals

We now consider the setting where only X has been recorded. Decrouez and Hall (2014) did not intend the split sample method to be used if only X is known, since it relies on the entire sample $X_1, \ldots, X_n$. It is however possible to apply their methodology also in this setting by constructing a random sequence $X_1, \ldots, X_n$ of 0's and 1's such that $\sum_{i=1}^n X_i = X$. We will refer to this interval as the externally randomized split sample interval, to distinguish it from Decrouez and Hall's original proposal. The externally randomized split sample interval uses a random noise term Y to improve the normal approximation. The next step is to ask whether there are other distributions for the noise that yield an even better approximation and Q2 thereby decrease the coverage oscillations even further. The answer is yes, for instance if X is replaced by $\dot{X} = X + Y$ where Q3 $Y \sim U(-1/2, 1/2)$ independently of X. The normal approximations of X, $\tilde{X}$ and $\dot{X}$ are compared in Fig. 2. We will refer to the interval based on $\dot{X}$ as the $U(-1/2, 1/2)$ interval.

A randomized confidence interval that does not rely on a normal approximation is the Stevens (1950) interval. It belongs to an important class of confidence intervals for p, consisting of intervals (p_L, p_U) where the lower bound p_L is such that

$$\nu_1 \cdot \binom{n}{X} p_L^X (1-p_L)^{n-X} + \sum_{k=X+1}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha/2$$

Please cite this article in press as: Thulin, M., On split sample and randomized confidence intervals for binomial proportions. Statistics and Probability Letters (2014), http://dx.doi.org/10.1016/j.spl.2014.05.005