Confidence interval for Binomial Proportions

Smruti Abhyankar and Gursharn Kaur

October 17, 2010

Abstract

In the Project 'Confidence interval for Binomial Proportios', we deal with the problem of obtaining the confidence interval for parameter p of *Binomial distribution.* We have used *the Normal approximation to Binomial, Bootstrap-t, Boostrap percentile* to fulfill this purpose. We have done this for different values of n and p and compared these confidence intervals on the basis of length of the confidence interval and observed coverage probability.

1 Introduction

We can obtain exact as well as approximate confidence interval for parameter p of Binomial Distribution. As Binomial distribution is a descrete one, exact confidence intervals are *conservative* i.e. actual confidence coefficient is greater than the nominal value and results into wider confidence intervals than those obtained by using approximations. Approximate confidence intervals can be easily calculated while exact confidence intervals can be obtained if binomial-table values are available. Hence many a times we prefer some good approximations. For some choices of n and p Normal approximation may not be a good approach and in such cases we can use Bootstrap confidence intervals.

1.1 What is confidence interval?

Confidence interval is an interval estimate which signifies a range within which the parameter is estimated to lie. Confidence interval is a measure of uncertainty involved in estimating unknown population parameter by using sample information.

2 Methodology

2.1 Normal Approximation

Let $X \sim Binom(n,p)$. In order to obtain a confidence interval for p we draw a randpm sample of size n from $Binom(1,p)$ say $X_1, X_2, ..., X_n$ and calculate the sample proportion $\hat{p}$ as it is an unbiased eatimate of p which is nothing but sample mean and from the Large sample theory we know that

$$n\bar{X} \sim AN(np, np(1-p))$$

If $\hat{p}$ is a sample proportion then

$$\hat{p} \sim AN\bigg(p, \frac{p(1-p)}{n}\bigg)$$

So, 95% confidence interval for p is $\bigg(\hat{p} - z_{0.025}\sqrt{var(\hat{p})}, \hat{p} + z_{0.025}\sqrt{var(\hat{p})}\bigg)$
Where $z_{0.025}$ is such that $P(Z \geq z_{0.025}) = 0.025$, where Z $\sim N(0,1)$

2.2 Bootstrap

It is a resampling plan and was introduced by Efron in 1979.
It adresses the question of sampling distribution. In the Bootsrap world the original sample $X_1, \cdots, X_n$ represents the population and $X_1^*, \cdots, X_n^*$ is a random boostrap sample drawn with replacement. The bootstrap idea is to estimate a desired property of a statistic by using an estimate of F i.e. distribution function of the original sample. Boostrapping based on empirical distribution function is called the nonparametric boostrap and that based on parametric estimate of F is called the parametric boostrap. Since we know the parametric form of the original sample from which the sample is drawn, we apply parametric boostrap.

- **Bootstrap-t**: Given n and p, we draw a random sample of size n from $Binom(1,p)$ say $X_1, X_2, ..., X_n$ and obtain a sample proportion $\hat{p}$.We draw B bootstrap samples and obtain Bootsrtap pivot as

$$Z^* = \frac{\hat{p^*} - \hat{p}}{\sqrt{var(\hat{p^*})}}$$

Where $\hat{p^*}$ is the bootstrap estimate of p and $var(\hat{p^*})$ is the bootstrap variance estimate of the parameter p. In order to obtain $var(\hat{p^*})$ generally we need nested bootsrtap smaples. Since statictic under consideration is sample mean, the exact variance can be calculated and hence nested bootstrap samples are not required. The exact expression for the variance is

$$\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{\hat{p}(1-\hat{p})}{n}$$

Once we obtain $Z_1^*, \cdots, Z_B^*$, Let $\hat{G_{Z^*}}$ be the ecdf of $Z_1^*, \cdots, Z_B^*$. Then boostrap-t confidence interval is

$$(\hat{p} - G_{Z^*}^{-1}(0.025)\sqrt{var(\hat{p})}, \hat{p} - G_{Z^*}^{-1}(0.975)\sqrt{var(\hat{p})})$$

- **Bootstrap-percentile** : We gnerate B bootstrap samples and compute sample proportion for each bootstrap sample say $p_1^*, \cdots, p_B^*$. Let $\hat{G}_{p^*}$ be its ecdf. Then bootstrap percentile confidence interval is

$$[\hat{G^{-1}}_{p^*}(0.025), \hat{G^{-1}}_{p^*}(0.975)]$$

2.3 Comparison

Following are the two criteria based on which we have compared confidence intervals obtained by applying these three methods.

- **Coverage Probability** : It is the probability that the true value of the parameter lies in the confidence interval. Coverage probability based on single confidence interval is either 0 or 1 which may lead to underestimation or overestimation of the true coverage probability. Hence we obtain N, say,confidence intervals. The proportion of confidence intervals which include the true value of parameter gives the coverage probability.

- **Length of CI** : The lower and upper boundaries of a confidence interval are confidence limits. They define the range of a confidence interval. The length of the confidence interval is nothing but the difference between the upper and lower confidence limits. It is obvious that the confidence intervals with smaller lengths are good interval estimates. Hence, while comparing the methods in terms of the length of the confidence interval, the method that results into comparatively smaller length confidence intervals is considered as more efficient.

3 Observation / Results

3.1 Comparison on the basis of coverage probability

The following tables give the coverage probability for different values of n and p on the basis of 5000 confidence intervals.

1. for $p = 0.05$

n	10	20	50	100	500	1000
Normal	0.420	0.634	0.909	0.875	0.938	0.942
Boot-t	0.385	0.638	0.890	0.954	0.939	0.943
Boot-p	0.374	0.636	0.911	0.931	0.938	0.940

2. for $p = 0.1$

n	10	20	50	100	500	1000
Normal	0.6518	0.8748	0.8798	0.9344	0.9452	0.9552
Boot-t	0.6290	0.8540	0.9690	0.9440	0.9324	0.9312
Boot-p	0.6406	0.8810	0.9296	0.9370	0.9382	0.9328

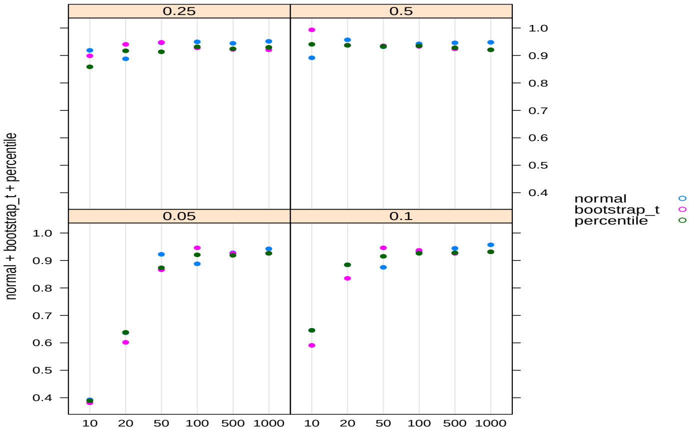
3. for $p = 0.25$

n	10	20	50	100	500	1000
Normal	0.9236	0.90827	0.9400	0.9448	0.9402	0.9438
Boot-t	0.9002	0.9396	0.9456	0.9344	0.9188	0.9194
Boot-p	0.8716	0.9332	0.9028	0.9294	0.9256	0.9216

4. for $p = 0.5$

n	10	20	50	100	500	1000
Normal	0.8874	0.9562	0.9344	0.9464	0.9452	0.9450
Boot-t	0.9930	0.9378	0.9342	0.9324	0.9320	0.9278
Boot-p	0.9382	0.9316	0.9366	0.9290	0.9274	0.9234

Dotplot corresponding to the above tables:



Conclusions:

1. For large n i.e. $n \geq 50$ coverage probability of confidence intervals obatained by using Normal approximation to Binomial and Bootstrap-t are almost same.

2. In case of small sample sizes, confidence intervals based on bootstrap methods are comparatively better than those obtained using Normal approximation to Binomial .

We can also conclude about the coverage probability of a given method with the help of following grpahs.
Comparison of the coverage probability of these three methods for chosen values $n = 100, p = 0.5, B = 100$.

- **for Normal**

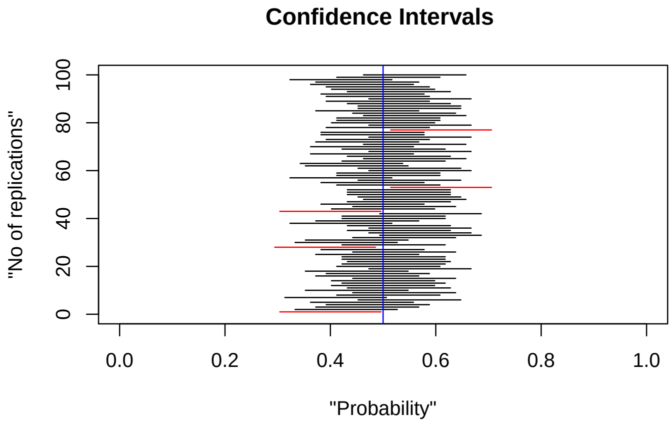


Figure 1: Confidence intervals using Normal approximation

- **for Bootstrap-t**

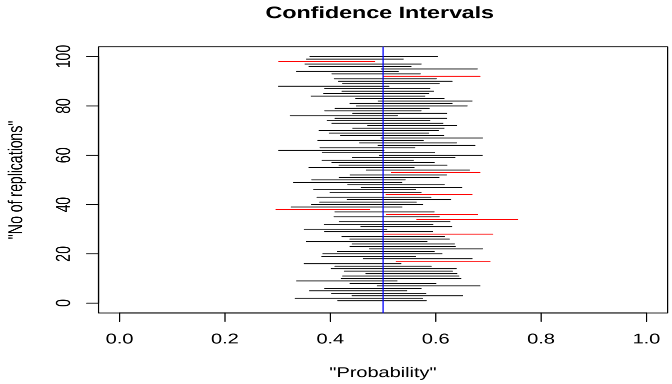


Figure 2: Boostrap-t confidence intervals

- **for Bootstrap-p**

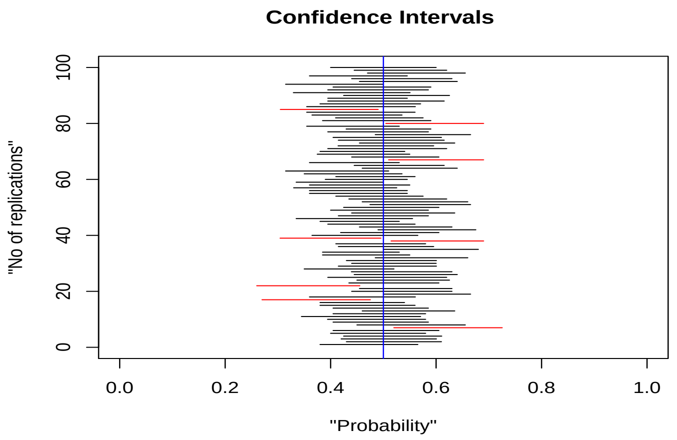


Figure 3: Confidence intervals using bootstrap-percentile

3.2 Comparison on the basis of length of CI

The following tables give the average length of the 5000 confidence intervals.

1. for $p = 0.05$

n	10	20	50	100	500	1000
Normal	0.16	0.144	0.112	0.083	0.038	0.0269
Boot-t	Inf	Inf	Inf	Inf	0.041	0.0278
Boot-p	0.1225	0.114	0.0967	0.0757	0.0366	0.0261

2. for $p = 0.1$

n	10	20	50	100	500	1000
Normal	0.289	0.235	0.161	0.116	0.052	0.0371
Boot-t	Inf	Inf	Inf	Inf	0.0542	0.0377
Boot-p	0.2303	0.199	0.1477	0.1103	0.051	0.0362

3. for $p = 0.25$

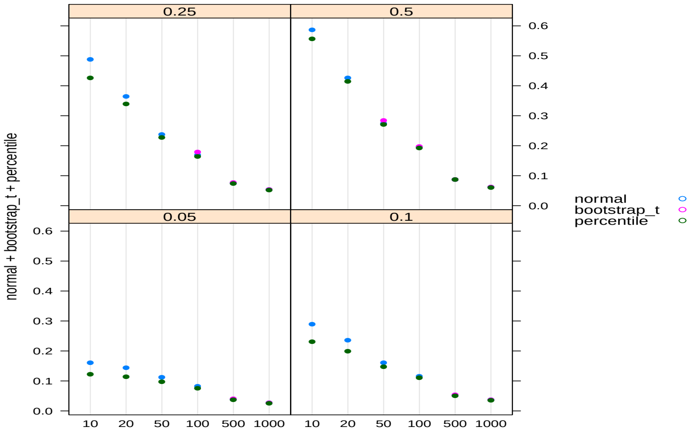
n	10	20	50	100	500	1000
Normal	0.487	0.365	0.237	0.169	0.0758	0.054
Boot-t	Inf	Inf	Inf	0.1793	0.0764	0.0535
Boot-p	0.4258	0.34	0.228	0.1648	0.0746	0.0527

4. for $p = 0.5$

n	10	20	50	100	500	1000
Normal	0.5868	0.4265	0.2744	0.195	0.0875	0.0619
Boot-t	Inf	Inf	0.2844	0.197	0.0870	0.0614
Boot-p	0.5569	0.414	0.2703	0.1927	0.086	0.0612

- **Conculsion-**In almost all the cases Bootstrap-percentile method provides shorter length confidence intervals as compared to the rest.

Dotplot corresponding to the above tables:



- **Conclusion -** All methods are almost equally efficient for large n.

4 Difficulties

When p is small, say $0.05, 0$ is more likely to occur when a sample of size n is drawn from Bernoulli distribution. It may happen that one or more bootstrap sample(s) consist(s) enitrely of zeroes. In such cases, sample proportion is zero and variance of sample proportion or sample mean i.e. statistic is zero. While calculating empirical distribution of critical values of boostrap confidence interval, this term is in the denominator. This leads to one sided confidence interval with infinite length. The same problem encounters when the parent sample from which further boostrap smaples are drawn, contains all zeroes.