Moved Score Confidence Intervals for Means of Discrete Distributions

Yu Guan

Department of Statistics, Zhejiang A & F University, Lin' an, Zhejiang, China
E-mail: guanyu@zafu.edu.cn

Abstract

Let *X* denote a discrete distribution as Poisson, binomial or negative binomial variable. The score confidence interval for the mean of *X* is obtained based on inverting the hypothesis test and the central limit theorem is discussed and recommended widely. But it has sharp downward spikes for small means. This paper proposes to move the score interval left a little (about 0.04 unit), called by moved score confidence interval. Numerical computation and Edgeworth expansion show that the moved score interval is analogous to the score interval completely and behaves better for moderate means; for small means the moved interval raises the infimum of the coverage probability and improves the sharp spikes significantly. Especially, it has unified explicit formulations to compute easily.

1. Introduction

Forming a confidence interval (*CI*) for the mean of a discrete distribution is one of the most basic problems in statistics, since the discrete lattice nature and skewness make the problem complicated. Let *X* be a discrete variable with the mean $E(X) = \mu$ and the variance $Var(X) = a\mu + b\mu^2$. For $X \sim \pi(\lambda)$ Poisson distribution with mean λ, $a = 1$, $b = 0$; for $X \sim B(n, p)$ binomial distribution with mean np, $a = 1$, $b = -1/n$; for $X \sim NB(r,p)$ negative binomial distribution, mean rp/q, $a = 1$, $b = 1/r$. Where $q = 1 - p$.

It is well known that the normal distribution $N(\mu, a\mu + b\mu^2)$ could be regarded as an approximation to *X* for large sample by the central limit theorem. Let

$$|X - \mu| / (a\mu + b\mu^2)^{1/2} \leq z, \text{ where } z = \Phi^{-1}(1 - \alpha/2) \text{ is}$$

the $100(1 - \alpha/2)$ th percentile of the standard normal distribution. We have the score confidence interval for the mean μ as follows

$$CI_S = \frac{X + 2^{-1}az^2}{1 - bz^2} \pm \frac{z}{1 - bz^2}\sqrt{aX + bX^2 + 4^{-1}a^2z^2} .$$

The score interval is an approximate interval and has many better properties for moderate mean μ. Those articles concern approximate confidence interval almost refer to the score interval. See references in this paper.

The exact confidence interval for the mean and the confidence level $1 - \alpha$ is obtained by solving equations $P\{X \geq x | \mu = L(x) = a_1\}$ and $P\{X \leq x | \mu = U(x) = a_2\}$, where $\alpha_1 + \alpha_2 = \alpha$.

In this paper, the exact confidence interval indicates the shortest length interval with less coverage probability not less than the nominal level (see [1]). Obviously, the exact interval has not explicit formulation and its computation will be troublesome if one does not use a computer. In general, an approximate interval cannot guarantee its coverage probabilities all not less than the nominal level, but its formula is simple and easily computed [2-5].

This paper discusses a moved score interval *CI(c)* for the means of Poisson, binomial and negative binomial variables. *CI(c)* with its coverage probability is introduced in section 2. From section 3 to section 4, *CI(c)* for the means of Poisson, binomial, negative binomial variables are discussed respectively. In section 5, Edgeworth expansion on coverage probabilities of *CI(c)* are investigated and compared. Conclusion and recommendation appears in the last section.

2. Moved Score Confidence Interval

Let *X* be a variable with the mean μ and the variance

$a\mu + b\mu^2$. A moved score confidence interval for the mean μ is defined as

$$CI(c) = \frac{X + 2^{-1}az^2}{1 - bz^2} \pm \frac{z}{1 - bz^2}\sqrt{aX + bX^2 + 4^{-1}a^2z^2}$$

Obviously, $CI(c)$ is equivalent to moving $CI(1/2) = CI_S$ left $1/2 - c$ units.

Theorem 1 The coverage probability of $CI(c)$ can be computed as

$$P\{\mu \in CI(c)\} = P\left\{\left|X - \frac{(1 - 2bcz^2)\mu + \left(\frac{1}{2} - c\right)az^2}{1 - bz^2}\right| \leq \right.$$
$$\left. \frac{(1 - 2bcz^2)z}{1 - bz^2}\sqrt{a\mu + b\mu^2 + \frac{\left(\frac{1}{2} - c\right) - \left(\frac{1}{4} - c^2\right)bz^2}{(1 - 2bcz^2)^2}a^2z^2}\right\}$$

Proof Let

$\mu \in \left[X + acz^2 \pm z\sqrt{X + 4^{-1}z^2}\right]\big/(1 - 2bcz^2)$. By some trivial deduction, the formula is obtained easily.

Following theorem is the specific case of theorem 1.

Theorem 2 1) If $X \sim \pi(\lambda)$, the moved score interval on the mean is defined as

$$CI_\pi(c) = X + cz^2 \pm z\sqrt{X + 4^{-1}z^2}$$

and its coverage probability is equal to

$$P\{\lambda \in CI_\pi(c)\}$$
$$= P\left\{\left|X - \lambda + (1/2 - c)z^2\right| \leq z\sqrt{\lambda + (1/2 - c)z^2}\right\}.$$

2) If $X \sim B(n, p)$, the moved score interval on the mean is defined as

$$CI_B(c) = \left[X + cz^2 \pm z\sqrt{X - n^{-1}X^2 + 4^{-1}z^2}\right]\big/(1 + 2cz^2n^{-1})$$

and its coverage probability is equal to

$$P\{np \in CI_B(c)\} = P\left\{\left|X - \frac{(n + 2cz^2)np + \left(\frac{1}{2} - c\right)nz^2}{n + z^2}\right| \right.$$
$$\left. \leq \frac{(n + 2cz^2)z}{n + z^2}\sqrt{npq + \frac{\left(\frac{1}{2} - c\right)n + \left(\frac{1}{4} - c^2\right)z^2}{(n + 2cz^2)^2\big/n}z^2}\right\}$$

3) If $X \sim NB(r, p)$, the moved score interval on the Mean is defined as

$$CI_{NB}(c)$$
$$= \left[X + cz^2 \pm z\sqrt{X + r^{-1}X^2 + 4^{-1}z^2}\right]\big/(1 - 2cz^2r^{-1})$$

and its coverage probability is equal to

$$P\left\{\frac{np}{q} \in CI_{NB}(c)\right\} = P\left\{\left|X - \frac{(r - 2cz^2)\frac{rp}{q} + \left(\frac{1}{2} - c\right)rz^2}{r - z^2}\right| \right.$$
$$\left. \leq \frac{(r - 2cz^2)z}{r - z^2}\sqrt{\frac{rp}{q^2} + \frac{\left(\frac{1}{2} - c\right)r - \left(\frac{1}{4} - c^2\right)z^2}{(r - 2cz^2)^2\big/r}z^2}\right\}$$

3. Moved Score Interval for the Mean of a Poisson Variable

Let $X \sim \pi(\lambda)$, $\lambda_L = k + z^2/2 - z(k + z^2/4)^{1/2}$. Set $k = 1, 2, 3$, $1 - \alpha = 0.95$ and $z = 1.95996$, we have $\lambda_L = 0.17652$, 0.54847, 1.02027 respectively. If λ take values less than λ_L a little as $\lambda = 0.1765$, 0.5484, 1.0202, small coverage probabilities arise.
$P\{0.1765 \in CI_\pi(1/2)\} = P\{X = 0 | \lambda = 0.1765\} = 0.8382$. 0.8382 is less than the nominal level 0.95 markedly. In the same way, $P\{0.5484 \in CI_\pi(1/2)\} = 0.8948$, $P\{1.0202 \in CI_\pi(1/2)\} = 0.9159$.

Let the mean λ be small near 0, its lower bound of confidence interval should be 0. That is to say, two-sided confidence interval is exactly one-sided interval for the small means. Denote λ_α^* as the upper bound on the mean that two-sided confidence interval can be replaced by one-sided interval, and it could be estimated approximately by $P\{X \geq 1 | \lambda_\alpha^*\} = \alpha$.

Set

$$c_{\alpha,k} = \frac{1}{2} - \frac{x + 2^{-1}z^2 - z\sqrt{x + 4^{-1}z^2} - \lambda_{\alpha,k}}{z^2}$$

where $\lambda_{\alpha,k}$ is be the real number satisfying $P\{X \geq k \mid \lambda = \lambda_{\alpha,k}\} = 1 - \alpha$; $k = 1, 2, \cdots, K_\alpha$; K_α is the largest integer such that $\lambda_{\alpha,k} \leq \lambda_\alpha^*$.

Figure 1 shows that most of coverage probabilities of $CI_\pi(0.45)$ and $CI_\pi(0.44)$ are not less than the level for $\lambda \leq \lambda_\alpha^*$. But they seem to be conservative. Numerical computation shows $c = 0.46$ is almost the best choice on Poisson, binomial and negative binomial variables for general levels 0.90, 0.95, 0.99. In the latter part of this paper would mainly demonstrate advantages of $CI(0.46)$ (see **Figure 2**).

An important criterion to judge a confidence interval
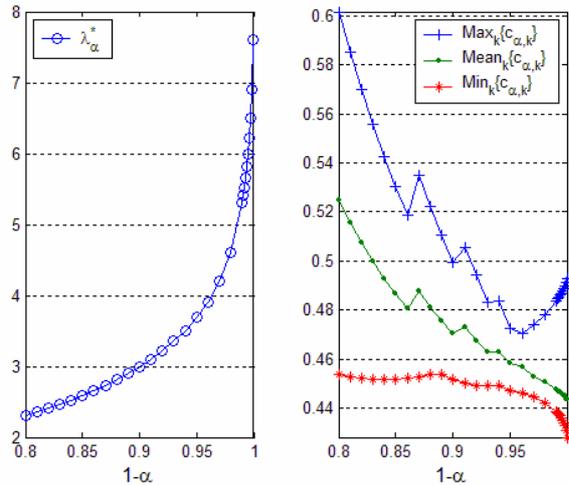
Figure 1. For $1 - \alpha = 0.80, 0.81, \cdots, 0.99$ and $1 - \alpha = 0.991, 00.992, \cdots, 0.999$, the left panel figures λ_a^* such that $P\{X \geq 1 | \lambda_a^*\} = \alpha/2$, and the right panel figures the maximum $\text{Max}_k\{c_{\alpha,k}\}$, the average $\text{Mean}_k\{c_{\alpha,k}\}$, and the minimum $\text{Min}_k\{c_{\alpha,k}\}$ for $k \leq K_\alpha$ from top to bottom curve.
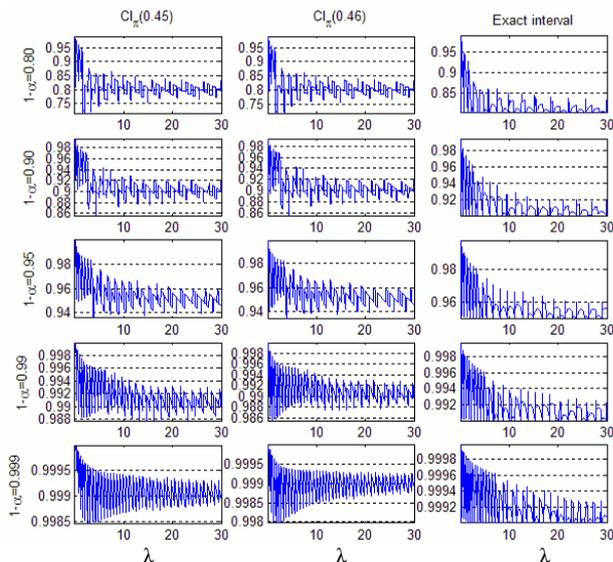
Figure 2. Coverage probabilities of the moved intervals $CI_\pi(0.45)$ (left panels), $CI_\pi(0.46)$ (middle panels) and exact interval (right panels) for $\lambda \in [0.01, 30]$ on $\pi(\lambda)$ with the levels 0.8, 0.90, 0.95, 0.99, 0.999 (from the top to the bottom panels).

is the confidence coefficient, *i.e.* the infimum of the coverage probability (*ICP*) of the interval [4]. If *ICP* < $1 - \alpha$, the larger is *ICP* the better is the interval.

Figure 2 and **Table 1** show that $CI_\pi(0.46)$ and $CI_\pi(0.45)$ greatly increase *ICP* and evidently improve the spike characteristic of $CI_\pi(0.5)$ for small λ. Of course, $CI_\pi(0.46)$ is more excellent than $CI_\pi(0.45)$.

Table 2 from [5] lists other eight approximate intervals as the First Normal, with Correction, Bartlett, Mole-

Table 1. The confidence coefficients of $CI_\pi(0.5)$, $CI_\pi(0.46)$, $CI_\pi(0.45)$ on Poisson variable.

$1 - \alpha$	0.85	0.90	0.95	0.99	0.995
$CI_\pi(0.5)$	0.7695	0.8001	0.8382	0.8896	0.9030
$CI_\pi(0.46)$	0.7976	0.8670	0.9344	0.9853	0.9910
$CI_\pi(0.45)$	0.7982	0.8558	0.9357	0.9877	0.9936

Table 2. The confidence coefficients of some other approximate confidence intervals on Poisson variable when $1 - \alpha = 0.95$.

Method	ICP	Method	ICP
First Normal	0.0769	with Correction	0.8798
Bartlett	0.6251	Molenaar	0.8458
Freeman and Tukey	0.9155	Anscombe	0.9101
midP	0.9165	Hald	0.5630

naar, Freeman and Tukey, Anscombe, midP, Hald interval, their *ICP* are equal to 0.0769, 0.8798, 0.6251, 0.8458, 0.9155, 0.9101, 0.9165, 0.5630 respectively. They are all worse than the moved score interval $CI_\pi(0.46)$.

For confidence levels $1 - \alpha$, **Table 3** lists ratios of the confidence probabilities located in intervals $[1 - a \pm a/5]$ on $CI_\pi(0.5)$, $CI_\pi(0.46)$, $CI_\pi(0.45)$ and the exact interval for $\lambda = 0.001t (= 1, 2, \cdots, 30000)$. When $1 - \alpha = 0.85$, 0.90, 0.95, 0.99, 0.995, intervals $[1 - a \pm a/5]$ = [0.82, 0.88], [0.88, 0.92], [0.94, 0.96], [0.988, 0.992], [0.994, 0.996] respectively. The larger is the ratio, the more is there coverage probabilities close to the level. **Table 3** shows that errors between coverage probabilities and levels on the score interval and moved score intervals are analogous to the exact interval, although they do not gurantee all coverage probabilities are not less than levels as the later.

## 4. Moved Score Intervals for the Means of Binomial Variable and Negative Binomial Variable

Let $X \sim B(n, p)$, when p near 0 and 1, $\pi(np)$ could be as an approximation to $B(n, p)$. So we pay attention to $CI_B(0.45)$ and $CI_B(0.46)$ also. Agresti and Coull [6], Agresti and Caffo [1] suggested the score interval and the Agresti-Coull interval; Brown *et al.* [3,7,8] recommended the Agresti-Coull interval, the modified Wilson (score) interval, modified Jeffreys interval and the likelihood ratio interval; Vollset [9] also recommended score methods for its easily computation; Zhou *et al.* [10] recommended the score interval if there is no available information about p. We believe that the score method is the uppermost approximation on interval estimation of a binomial proportion.

Table 3. Ratios of confidence probabilities of $CI_\pi(0.5)$, $CI_\pi(0.46)$, $CI_\pi(0.45)$ and the exact interval on Poisson variable.

$1-\alpha$	0.85	0.90	0.95	0.99	0.995
$CI_\pi(0.5)$	0.8402	0.8498	0.7838	0.7094	0.5915
$CI_\pi(0.46)$	0.8336	0.8205	0.7609	0.6589	0.6440
$CI_\pi(0.45)$	0.8316	0.8136	0.7525	0.6290	0.5937
CI_{Exact}	0.8765	0.8018	0.8239	0.6821	0.5570

Figure 3 shows that intervals $CI_B(0.46)$ and $CI_B(0.45)$ improves the spikes of $CI_B(0.5)$ obviously. For small p the Agresti-Coull interval behaves too conservative. The Jeffreys interval is a better interval for moderate p, but it has sharp spikes for small p also. Brown *et al.* [3] suggested revising two specific limits when $X = 0, 1, n-1$, or n. Besides, they used one-sided Poisson approximation to binomial distribution to modify CI_S with $X = 1,2$ for $n < 50$ and $X = 3$ for $n \geq 50$. Numerical computation shows the modified score interval and the modified Jeffreys interval are comparable with moved score intervals $CI_B(0.46)$, but the latter method and formula are more simple than the formers. Zhou *et al.* [10] proposed ZL interval based on logit transformation, but its coverage probabilities are greater than the nominal level when p is close to 0 or 1.

Let $X \sim NB(r,p)$, when p near 0 and r large, $\pi(rp/q)$ could be as an approximation to $NB(r, p)$. By numerical computation, we believe $CI_{NB}(0.45)$ and $CI_{NB}(0.46)$ improve the spikes of $CI_{NB}(0.5)$ obviously also. There is
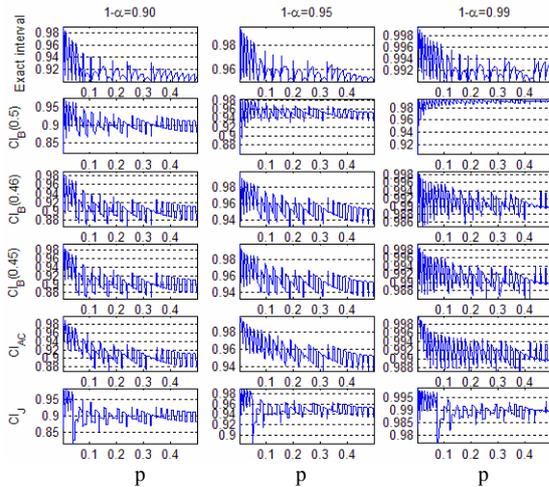
Figure 3 Coverage probabilities of the exact interval (the first row panel), the score interval $CI_B(0.5)$ (the second row panel), the moved intervals $CI_B(0.46)$ (the third row panel) and $CI_B(0.45)$ (the forth row panel), the Agresti-Coull interval CI_{AC} (the fifth row panel) and the Jeffreys interval CI_J (the bottom panel) for $B(50, p)$ with $p = 0.001, 0.002,\cdots,0.500$ for levels 0.90, 0.95, 0.99 (from the left to the right panels).

fewer people interesting confidence interval on negative binomial variable than binomial variable markedly.

5. Edgeworth Expansion

Brown *et al.* [3,7,8] suggested utilizing Edgeworth expansion to theoretically analyze the coverage probability of a confidence interval. In general, the intervals for Poisson, binomial and negative binomial variable based on the same method almost have the same Edgeworth expansion (see [8]). So, we only discuss Edgeworth expansion of the moved interval on binomial variable in this section.

Let $P(np \in CI_B(c)) = P\left(l_c \leq \dfrac{X/n - p}{\sqrt{pq/n}} \leq u_c \right),$

where

$$[l_c, u_c] = \frac{(1-2c)(1-2p)z^2}{2(n+z^2)\sqrt{pq/n}}$$

$$= \pm\frac{(n+2cz^2)z}{n+z^2}\sqrt{1+\frac{\left(\frac{1}{2}-c\right)n+\left(\frac{1}{4}-c^2\right)z^2}{(n+2cz^2)^2 \, pq}z^2}$$

$$= \pm z + \frac{(1-2c)(1-2p)z^2}{2\sqrt{npq}} \pm \frac{(1-2c)(1-4pq)z^3}{4npq}$$

$$+ O\left(n^{-3/2}\right)$$

Set $\lambda_1 = \pm z$, $\lambda_2 = \dfrac{(1-2c)(1-2p)z^2}{2\sqrt{pq}}$,

$\lambda_3 = \dfrac{(1-2c)(1-4pq)z^3}{4pq}$, By lemma 1 in [3], we obtain

Theorem 3 Suppose $np - z\sqrt{npq}$ is not an integer. Then the coverage probability of $CI_B(c)$ satisfies

$$P(np \in CI_B(c)) = 1 - \alpha + \left[g(p, l_c) - g(p, u_c) \right]\frac{\varphi(z)}{\sqrt{npq}}$$

$$+ \left[\frac{(6c-1)(1-2c)(1-2p)^2 z^5}{12pq} + W_S \right]\frac{\varphi(z)}{n}$$

$$+ \left[(1-2p)\left(\frac{z^2}{6}-\frac{1}{2}\right)Q_{21}(l_c,u_c) + Q_{22}(l_c,u_c) \right]z\frac{\varphi(z)}{npq}$$

$$+ O\left(n^{-3/2}\right)$$

where

$$W_S = \left(\frac{1}{9}-\frac{1}{36pq}\right)z^5 + \left(\frac{7}{36pq}-\frac{11}{18}\right)z^3 + \left(\frac{1}{6}-\frac{1}{6pq}\right)$$

W_S is the coefficient of $O(n^{-1})$ nonoscillating term of the score interval $CI_B(0.5)$.

Remark 1. In theorem 3, the first $O(n^{-1})$ term is nonoscillating and would produce systematic bias without it. So it is a key term. We called its coefficient by coefficient of $O(n^{-1})$ nonoscillating term. Meanings of other terms in theorem 3 are explained in detail in [3,8].

By theorem 3, the coefficient of $O(n^{-1})$ nonoscillating term of $CI(c)$ is

$$W_{MS}(c) = \frac{(6c-1)(1-2c)(1-2p)^2 z^5}{12pq} + W_S$$

Coefficients of $O(n^{-1})$ nonoscillating terms of the Agresti-Coull interval, the Jeffreys interval and the likelihood ratio interval list as follows (from [8]):

$$W_{AC} = \left(\frac{1}{4pq} - 1\right) z^3$$

$$W_J = -\frac{1}{12pq} z$$

$$W_{LR} = \left(\frac{1}{6} - \frac{1}{6pq}\right) z$$

It is easily validated that
$W_{AC} \geq W_{MS}(0.46) \geq W_S \geq W_J \geq W_{LR}$.

See **Figure 4**. For moderate p, the coefficient of $CI_B(0.5)$ is the most close to 0. This ensures the score interval behaves well for moderate p. The moved intrval $CI_B(0.46)$ is more conservative than $CI_B(0.5)$ a little. The CI_{AC} is too conservative obviously. Coefficients of CI_J and CI_{LR} are not larger than $-z/3$ and $-z/2$, and too small. Let $n = 200$, by computation, ratios of coverage probabilities of CI_J not larger than levels 0.90, 0.95 and 0.99 are about 53%, 51% and 54% respectively. In the meanwhile, ratios of $CI_B(0.46)$ and CI_{AC} are about 48%, 43%, 40% and 42%, 34%, 28% respectively. Thus, the Jeffreys interval is more stingy than $CI_B(0.46)$.
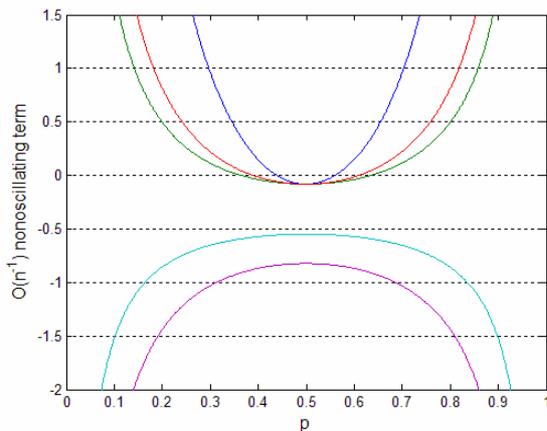
Figure 4. Coefficients of $O(n^{-1})$ nonoscillating terms of the coverage probabilities of CI_{AC}, $CI_B(0.46)$, $CI_B(0.5)$, CI_J and CI_{LR} (from the top to the bottom curve) with levels $1 - \alpha = 0.80, 0.90, 0.95, 0.99$.

Brown *et al.* [5] showed the ranking expected length of the intervals is CI_{LR}, CI_J and $CI_B(0.5)$ from the shortest to the longest, provided $z > 0.86$. Of course their differences are less. On the other hand, the length of interval $CI_B(0.46)$ is equal to the score interval for moderate p. Therefore, the expected length of $CI_B(0.46)$ is larger than CI_{LR} and CI_J a little for $z > 0.86$.

6. Conclusions and Recommendation

The score interval is concerned all the time by many statisticians for simple formula and good natures. But it has sharp downward spikes for the small mean, since discreteness and skewness cause this problem. Moving the score interval left a little could improve it, though spike phenomena could not be overcome completely.

We recommend the moved score intervals $CI_\pi(0.46)$, $CI_B(0.46)$ and $CI_{NB}(0.46)$ respectively for the means of Poisson variable, binomial variable and negative binomial variable as follows

$$CI_\pi(0.46) = X + 0.46z^2 \pm z\sqrt{X + 4^{-1}z^2}$$

$$CI_B(0.46) = \frac{X + 0.46z^2}{1 + 0.92\, n^{-1}z^2} \pm z\frac{\sqrt{X(1 - X/n) + 4^{-1}z^2}}{1 + 0.92\, n^{-1}z^2}$$

$$CI_{NB}(0.46) = \frac{X + 0.46z^2}{1 - 0.92\, r^{-1}z^2} \pm z\frac{\sqrt{X(1 + X/r) + 4^{-1}z^2}}{1 - 0.92\, r^{-1}z^2}$$

Especially for small Means $CI(0.45)$ is analogous to $CI(0.46)$, but $CI(0.45)$ behaves more conservative than $CI(0.46)$.

7. Acknowledgements

8. References

[1] P. Kabaila and J. Byrne, "Exact Short Poisson Confidence Intervals," *Canadian Journal of Statistics*, Vol. 29, No. 1, 2001, pp. 99-106. doi:10.2307/3316053

[2] A. Agresti and B. Coull, "Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, Vol. 52, No. 2, 1998, pp. 119-126. doi:10.2307/2685469

[3] L. D. Brown, T. T. Cai and A. DasGupta, "Interval Estimation intervals for a Binomial Proportion and Asymptotic Expansion," *The Annals of Statistics*, Vol. 30, No. 1, 2002, pp. 160-201.

[4] G. Casella and R. L. Berger, "Statistical Inference," 2nd Edition, Wadsworth, West Yorkshire, 2002.

[5] J. Byrne and P. Kabaila, "Comparison of Poisson Confi-

dence Intervals," *Communications in Statistics-Theory and Methods*, Vol. 34, No. 3, 2005, pp. 545-556. doi:10.1081/STA-200052109

[6] A. Agresti and B. Caffo, "Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures," *The American Statistician*, Vol. 54, No. 4, 2000, pp. 280-288. doi:10.2307/2685779

[7] L. D. Brown, T. T. Cai and A. DasGupta, "Confidence Intervals for a Binomial Proportion and Asymptotic Expansion," *The Annals of Statistics*, Vol. 30, No. 1, 2002,

pp. 160-201.

[8] L. D. Brown, T. T. Cai and A. DasGupta, "Interval Estimation in Exponential Families," *Statistica Sinica*, Vol. 13, 2003, pp. 19-49.

[9] S. E. Vollset, "Confidence Intervals for a Binomial Proportion," *Statistics in Medicine*, Vol. 12, No. 9, 1993, pp. 809-824. doi:10.1002/sim.4780120902

[10] X. H. Zhou, C. M. Li and Z. Yang, "Improving Interval Estimation of Binomial Proportions," *Philosophical Transactions of the Royal Society A*, Vol. 366, No. 1874, 2008, pp. 2405-2418. doi:10.1098/rsta.2008.0037