Confidence Intervals for the Binomial Proportion: A Comparison of Four Methods

Luke Akong'o Orawo

Mathematics Department, Egerton University, Egerton, Kenya
Email: luke.orawo@egerton.ac.ke

Abstract

This paper presents four methods of constructing the confidence interval for the proportion p of the binomial distribution. Evidence in the literature indicates the standard Wald confidence interval for the binomial proportion is inaccurate, especially for extreme values of p. Even for moderately large sample sizes, the coverage probabilities of the Wald confidence interval prove to be erratic for extreme values of p. Three alternative confidence intervals, namely, Wilson confidence interval, Clopper-Pearson interval, and likelihood interval, are compared to the Wald confidence interval on the basis of coverage probability and expected length by means of simulation.

Keywords

Binomial Distribution, Confidence Interval, Coverage Probability, Expected Length, Relative Likelihood Function

1. Introduction

Estimation of a binomial proportion p is one of the most commonly encountered statistical problems, with important application in areas such as clinical medicine, business, politics and quality control. For instance, politicians are certainly interested in knowing what fraction of voters would favor them in the next election. The binomial data is obtained from a binomial experiment which consists of a fixed number n of independent Bernoulli trials, each of which can result in either a success or a failure. The success probability p is assumed fixed. The binomial probability distribution is used to model the total number x of success resulting from the Binomial experiment. Once data are available, then information about p can be summarized by the likelihood function and on the basis of this summary, a point estimate for the Binomial proportion, denoted by $\hat{p}$ is obtained by the method of maximum likelihood as $\hat{p} = \dfrac{x}{n}$. A number of

two-sided confidence intervals for p have been proposed by several authors. The Wald method is the most commonly used technique since it is based normal approximation to the binomial distribution. However, the approximation is inaccurate whenever the sample size is small ($n < 30$) or when the proportion p is close to zero or one; the Wald confidence interval may have low coverage probability even if p is not close to zero or one, and confidence limits outside the interval $(0,1)$. Matiri *et al.* [1] applied the Wald method to obtain interval estimates for the prevalence rate and encountered the problem of overshoot with negative lower confidence limits. Poor performance of the Wald confidence interval has been pointed out by many authors [2] [3] [4] [5] [6].

Alternative methods for constructing confidence interval for p have been proposed, such as the Wilson Score, Clopper-Pearson and Agresti-Coull confidence intervals among others. Just like the Wald confidence intervals, the validity of the Wilson Confidence interval heavily depends on large sample approximation. The Clopper-Pearson interval is an exact two-sided confidence interval derived from the binomial probability mass function. Past studies indicate that the Clopper-Pearson confidence interval is very conservative for small to moderate n [3]. Panatiogis & Konstantinos [7] present a bootstrap method for estimating the binomial proportion and compare it with Wald confidence interval and Agresti-Coull interval.

This paper considers an alternative method, called the likelihood method, for constructing the approximate confidence interval for the binomial proportion. The likelihood intervals are determined from the graph of the relative likelihood function or its logarithm for a fixed likelihood level [8]. They are fully conditioned on the shape of the likelihood function and hence are optimal. The likelihood method can be used to construct confidence interval for the proportion in situations where the traditional methods based on asymptotic normality are inaccurate.

In order to identify the best confidence interval for the binomial proportion *p*, the Wald, Wilson score, Pearson-Clopper and Likelihood methods of interval estimation are compared on the basis of coverage probability and interval width using simulated data. The four intervals are also applied to a real data example. The resulting confidence intervals for the binomial proportion are compared in terms interval width and plausibilities of the parameter values in them.

The paper is organized as follows: in Section 2, the four methods of interval estimation are described. In Section 3, the simulation results regarding coverage probability and expected length of the different intervals are presented and discussed. Section 4 applies the four intervals to a real-life data from a clinical study and compares them in terms of interval length and plausibilities of the parameter values inside them. Section 5 is devoted to concluding remarks.

2. Interval Methods

2.1. Wald Interval

Let $X_1, \cdots, X_n$ be IID Bernoulli (p) random variables, where the parameter

$p \in (0,1)$ is unknown. Then the sum $X = \sum_{i=1}^{n} X_i$ of the n Bernoulli random variables is a binomial random variable with parameters n and p. If the unknown proportion p is not too close to 0 or 1, then by the Central Limit Theorem, for n sufficiently large, the MLE $\hat{p} = \dfrac{X}{n}$ is approximately normally distributed with mean $\mu_{\hat{p}} = p$ and variance $\sigma_{\hat{p}}^2 = \dfrac{p(1-p)}{n}$. The Wald confidence interval is based on the normal approximation to the binomial distribution and is given by $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, where $z_{\frac{\alpha}{2}}$ is the $1 - \dfrac{\alpha}{2}$ percentile of the standard normal distribution. The Wald method should be used only when $n * \min(p, 1-p)$ is at least 5 (or 10), otherwise it will produce unreliable interval estimates.

2.2. Clopper-Pearson Interval

Clopper-Pearson [9] proposed a method of constructing an exact two-sided confidence interval for the binomial proportion p using the equal-tail rule. The derivation of the two-sided $100(1-\alpha)\%$ Clopper-Pearson confidence interval for the binomial proportion p is based on the relationships between the binomial, beta and F distributions. The relationships are stated in the following three theorems.

Theorem 1

If $X \sim Beta(\alpha, \beta)$ then $Z = 1 - X \sim Beta(\beta, \alpha)$

Proof

The density function of X is given by $f(x) = \dfrac{\Gamma(\alpha + \beta)}{\Gamma(x)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$. By change of variable technique the density function of Z is obtained as

$$f_z(z) = f_x(1-z)\left|\dfrac{dx}{dz}\right| = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(x)\Gamma(\beta)}(1-z)^{\alpha-1} z^{\beta-1},$$

which is the density function of a beta distribution with parameters β and α. Implying that $Z \sim Beta(\beta, \alpha)$.

Theorem 2

If $X \sim Bin(n, p)$ then $P_p[X \geq x] = P[Y \leq p]$, where $Y \sim Beta(x, n-x+1)$

Proof

Consider the identity

$$\sum_{k=0}^{x}\binom{n}{k} p^k (1-p)^{n-k} = (n-x)\binom{n}{x}\int_0^{1-p} t^{n-x-1}(1-t)^x \, dt, \qquad \text{(i)}$$

We use the above identity to obtain

$$P[X \geq x] = 1 - P[X \leq x-1]$$

$$= 1 - \sum_{k=0}^{x-1}\binom{n}{k} p^k (1-p)^{n-k}$$

$$= 1 - (n - x + 1)\binom{n}{x-1}\int_0^{1-p} t^{n-(x-1)-1}(1-t)^{x-1}\,\mathrm{d}t$$

$$= 1 - \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}\int_0^{1-p} t^{n-(x-1)-1}(1-t)^{x-1}\,\mathrm{d}t$$

$$= 1 - P[T \le 1-p],$$

where

$$T \sim Beta(n-x+1, x)$$

$$= P[T \ge 1-p]$$

$$= P[-T \le p-1]$$

$$= P[1-T \le p]$$

$$= P[Y \le p]$$

where $Y = 1 - T$.

Hence it follows by Theorem 1 that $Y \sim Beta(x, n-x+1)$.

Theorem 3

If X has an F distribution with u and v degrees of freedom, then the random

variable $Y = \dfrac{\dfrac{u}{v}X}{1 + \dfrac{u}{v}X}$ has a $Beta\left(\dfrac{u}{2}, \dfrac{v}{2}\right)$ distribution.

Proof

Let $y = \dfrac{\dfrac{u}{v}x}{1 + \dfrac{u}{v}x}$. Then $x = \dfrac{y}{1-y}\dfrac{v}{u}$ and $\dfrac{\mathrm{d}x}{\mathrm{d}y} = \dfrac{1}{(1-y)^2}\dfrac{v}{u}$. By the change of

variable technique the density function of Y is obtained as

$$f_Y(y) = f_X\left(\frac{y}{1-y}\frac{v}{u}\right)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$$

$$= \frac{\Gamma\left(\dfrac{u+v}{2}\right)\left(\dfrac{u}{v}\right)^{\frac{u}{2}}\left(\dfrac{y}{1-y}\dfrac{v}{u}\right)^{\frac{u}{2}-1}}{\Gamma\left(\dfrac{u}{2}\right)\Gamma\left(\dfrac{v}{2}\right)\left(1+\dfrac{y}{1-y}\right)^{\frac{u+v}{2}}}\frac{1}{(1-y)^2}\frac{v}{u}$$

$$= \frac{\Gamma\left(\dfrac{u+v}{2}\right)y^{\frac{u}{2}-1}(1-y)^{\frac{v}{2}-1}}{\Gamma\left(\dfrac{u}{2}\right)\Gamma\left(\dfrac{v}{2}\right)}$$

which is the density function of a $Beta\left(\dfrac{u}{2}, \dfrac{v}{2}\right)$ distribution. Hence

$Y \sim Beta\left(\dfrac{u}{2}, \dfrac{v}{2}\right)$.

The above three theorems are now applied in the derivation of the closed

forms of the lower and upper confidence limits of the Clopper-Pearson interval for the binomial proportion p as follows: Suppose that $Y \sim Beta(x, n-x+1)$, where x is the observed value of a $Bin(n, p)$ random variable X, then by Theorem 3 the random variable $\dfrac{n-x+1}{x}\dfrac{Y}{1-Y}$ has an F distribution with $2x$ and $2(n-x+1)$ degrees of freedom. Therefore for a fixed $\alpha \in (0,1)$, the lower limit of a two-sided exact Clopper-Pearson interval is obtained by solving the equation,

$$\frac{\alpha}{2} = P[X \geq x]$$

By Theorem 2 we have

$$\frac{\alpha}{2} = P[X \geq x] = P[Y \leq p]$$

where $Y \sim Beta(x, n-x+1)$

$$= P\left[\frac{n-x+1}{x}\frac{Y}{1-Y} \leq \frac{n-x+1}{X}\frac{p}{1-p}\right]$$

$$= P\left[F_{2x,2(n-x+1)} \leq \frac{n-x+1}{x}\frac{p}{1-p}\right],$$

where $F_{2x,2(n-x+1)}$ is an F random variable with $2x$ and $2(n-x+1)$ degrees of freedom. This implies that $f_{1-\frac{\alpha}{2},2x,2(n-x+1)} = \dfrac{n-x+1}{x}\dfrac{p}{1-p}$ and solving for p we get $\dfrac{1}{1+\dfrac{n-x+1}{x}f_{\frac{\alpha}{2},2(n-x+1),2x}}$ as the lower limit.

Similarly, the upper limit is obtained by solving the equation

$$\frac{\alpha}{2} = P[X \leq x]$$

Equivalently, we write

$$\frac{\alpha}{2} = P[X \leq x]$$

$$= 1 - P[X \geq x+1]$$

$$= 1 - P[T \leq 1-p],$$

where

$$T \sim Beta(n-x, x+1)$$

$$= P[T \geq 1-p]$$

$$= P[1-T \geq p]$$

$$= P[Y \geq p],$$

where

$$Y \sim Beta(x+1, n-x)$$

$$= P\left[\frac{n-x}{x+1}\frac{Y}{1-Y} \geq \frac{n-x}{x+1}\frac{p}{1-p}\right]$$

$$= P\left[F_{2(x+1),2(n-x)} \geq \frac{n-x}{x+1}\frac{p}{1-p}\right].$$

Solving this equation for p yields $\dfrac{\dfrac{x+1}{n-x}f_{\frac{\alpha}{2},2(x+1),2(n-x)}}{1+\dfrac{x+1}{n-x}f_{\frac{\alpha}{2},2(x+1),2(n-x)}}$ as the upper limit.

Therefore, the $100(1-\alpha)\%$ exact Clopper-Pearson confidence interval for p becomes

$$\frac{1}{1+\dfrac{n-x+1}{x}f_{\frac{\alpha}{2},2(n-x+1),2x}} \leq p \leq \frac{\dfrac{x+1}{n-x}f_{\frac{\alpha}{2},2(x+1),2(n-x)}}{1+\dfrac{x+1}{n-x}f_{\frac{\alpha}{2},2(x+1),2(n-x)}}.$$

2.3. Likelihood Interval

Let x be the observed value of a $Bin(n,p)$ random variable X. The likelihood function of p is defined as

$$L(p) = kP[X = x; p],$$

where k is any positive constant not depending on p. We choose k to simplify the expression for $L(p)$ and a natural choice is $k = \dfrac{1}{\dbinom{n}{x}}$. Then binomial likelihood function is

$$L(p) = p^x(1-p)^{n-x} \quad \text{for} \quad 0 < p < 1.$$

The log-likelihood function is now

$$l(p) = x\log(p) + (n-x)\log(1-p), \text{ for } 0 < p < 1.$$

The relative likelihood function of p, denoted by $R(p)$ is given by

$$R(p) = \frac{L(p)}{L(\hat{p})} = \frac{p^x(1-p)^{n-x}}{\left(\dfrac{x}{n}\right)^x\left(1-\dfrac{x}{n}\right)^{n-x}} = \left(\frac{np}{x}\right)^x\left(\frac{n(1-p)}{n-x}\right)^{n-x}$$

The log-relative likelihood function of p, denoted by $r(p)$ is

$$r(p) = \log R(p) = l(p) - l(\hat{p}) = x\log(p) + (n-x)\log(1-p) - l(\hat{p}).$$

The likelihood intervals may be determined from a graph of $R(p)$ or its logarithm, $r(p)$ although it is more convenient to work with $r(p)$. The set of p values for which $R(p) \geq c$ is called a $100c\%$ likelihood interval (LI). The maximum likelihood estimate (MLE) p, of $\hat{p}$ is the most plausible value of p in that it makes the observed sample most probable. The relative-likelihood function measures the plausibility of any specific value of p relative to that of $\hat{p}$. The end points of the $100c\%$ likelihood interval (LI) are obtained as the roots of

the equation $r(p) - \log(c) = 0$. The use of a numerical procedure is usually necessary to solve this equation. In repeated samples from the parent distribution $Bin(n, p)$ using arbitrary value of p, the resulting population of level c likelihood intervals will contain this value of p with known frequency. They are therefore also confidence intervals and so are likelihood confidence intervals.

2.4. Wilson Interval

The Wilson score method for constructing confidence interval for binomial proportion p was developed by Edward B. Wilson [10] and is based on inverting the z-test for p. The endpoints of the $100(1-\alpha)\%$ is obtained by solving the quadratic inequality $-z_{\frac{\alpha}{2}} \le \dfrac{\hat{p} - p}{\sqrt{pq/n}} \le z_{\frac{\alpha}{2}}$ for p. This confidence interval is of

the form $\dfrac{\left(2n\hat{p} + z_{\frac{\alpha}{2}}^2\right) \pm \sqrt{z_{\frac{\alpha}{2}}^2 + 4n\hat{p}\left(1 - \hat{p}\right)}}{2\left(n + z_{\frac{\alpha}{2}}^2\right)}$. The score confidence interval is

asymmetric and does not suffer from problems of overshoot and zero width confidence intervals associated with Wald confidence interval.

3. Simulations

In this section the simulation studies are carried out and finite-sample comparisons of the performances of the Wald, Cloper-Pearson, Wilson score and Likelihood intervals on the basis of coverage probability and expected length. For any confidence interval method for estimating of p, the actual coverage probability at a fixed value of p is

$$Cp(p, n) = \sum_{k=0}^{n} I(k, n) \binom{n}{k} p^k (1-p)^{n-k},$$

where $I(k, n)$ equals 1 if the interval contains p when $X = k$ and equals 0 if it does not contain p. Denote by $L(X)$ and $U(X)$ the lower and upper confidence limits, respectively. The expected length of this interval

$$EL(p, n) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \left[U(x) - L(x)\right]$$

The coverage probability and expected length were computed for 1000 values of p, equally spaced in the interval (0.2, 0.8) for sample sizes n = 15, 30, 50 and 100, and for nominal 95% Clopper-Pearson, Wilson score, Wald and likelihood confidence intervals. For each sample size and for each method summary values for coverage probability and expected length are obtained by averaging over the values of p used in the simulation. Table 1 below shows the mean of the actual coverage probabilities for the four methods of interval estimation at various sample sizes. The Clopper-Pearson interval is very conservative but has the highest mean interval length for all the values of n. The mean coverage probabilities

Table 1. Mean coverage probabilities and mean expected lengths (in parentheses) of nominal 95% confidence intervals for the binomial parameter p.

Method	$n = 15$	$n = 30$	$n = 50$	$n = 100$
Clopper-Pearson	0.974 (0.479)	0.969 (0.347)	0.965 (0.269)	0.961 (0.190)
Wilson score	0.957 (0.416)	0.951 (0.313)	0.950 (0.2487)	0.950 (0.180)
Wald	0.913 (0.456)	0.931 (0.328)	0.939 (0.256)	0.945 (0.182)
Likelihood	0.951(0.431)	0.948(0.319)	0.950 (0.216)	0.949 (0.81)

for Wilson interval are very close to the nominal level and has the smallest mean expected length for all n. On the other hand, the traditional Wald interval has mean coverage probabilities which are smaller than the nominal level. Finally, the mean coverage probabilities for likelihood interval are very close to the nominal level for all the sample sizes.

Figure 1 and Figure 2 show, respectively, plots of coverage probability and expected length against the values of p for the four intervals when $n = 15$. It can be noted from Figure 1 that for the Wald interval most coverage probabilities are below the nominal level and are extremely low for values of p near 0.2 or 0.8. This may be due to poor normality approximation when the sample is small and p not close to 0.5. Clopper-Interval has coverage probabilities above the nominal level and with short spikes, but has the largest expected lengths. The Wilson and likelihood intervals are not conservative but their coverage probabilities are close to the nominal level and have smaller expected lengths than Clopper-Pearson and Wald intervals (see Figure 2).

For a large sample $n = 50$, the same pattern is observed but there is a remarkable improvement in terms convergence of coverage probabilities and reduced expected lengths. Clopper-Pearson is still conservative and show convergence to a value above the nominal level. Most coverage probabilities for Wald interval are still below nominal level and show poor convergence. The Wilson and Likelihood interval again are better than Clopper-Pearson and Wald interval in terms of the two performance measures (Figure 3 and Figure 4).

4. Application to Real Example

The four methods of interval estimation are applied in a clinical study about the effectiveness of hyperdynamic therapy in treating cerebral vasospasm [11]. The success of the therapy was defined as clinical improvement in terms of neurological deficits. The study reported 16 successes out of 17 patients. On the basis of this data the four 95% confidence intervals are computed as 1) (0.7131, 0.9985) for the Clopper-Pearson interval, 2) (0.7302, 0.9895) for Wilson interval, (0.8289, 1.053) for Wald interval, and 3) (0.7658, 0.9965) for likelihood interval. Each of these four confidence intervals is plotted on the graph of relative likelihood function as shown in Figure 5. It is observed that the Clopper-Pearson and Wilson intervals include implausible values of the parameter p whereas the Wald interval excludes plausible values and has an upper limit greater than 1.
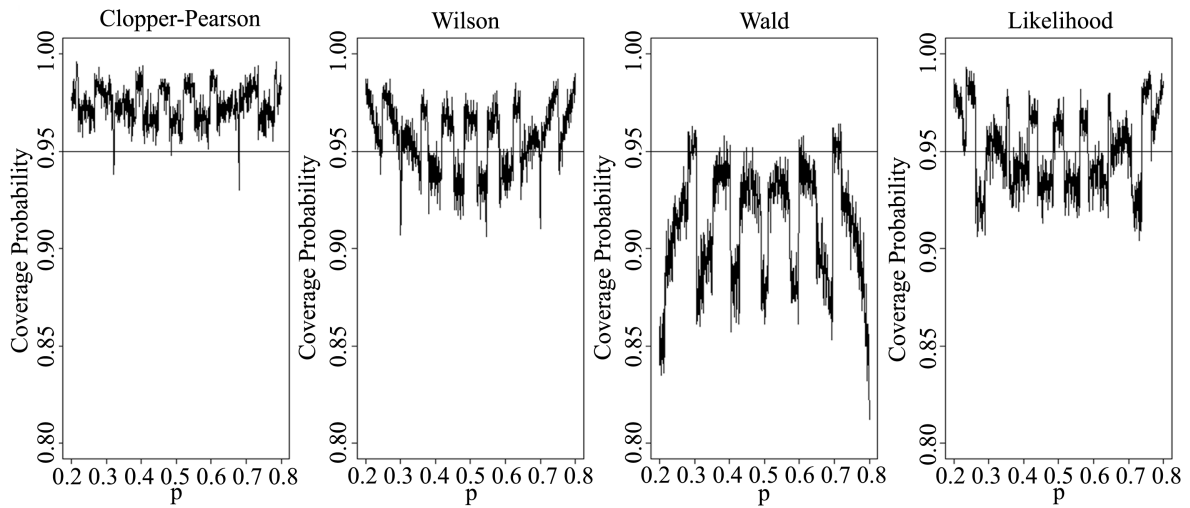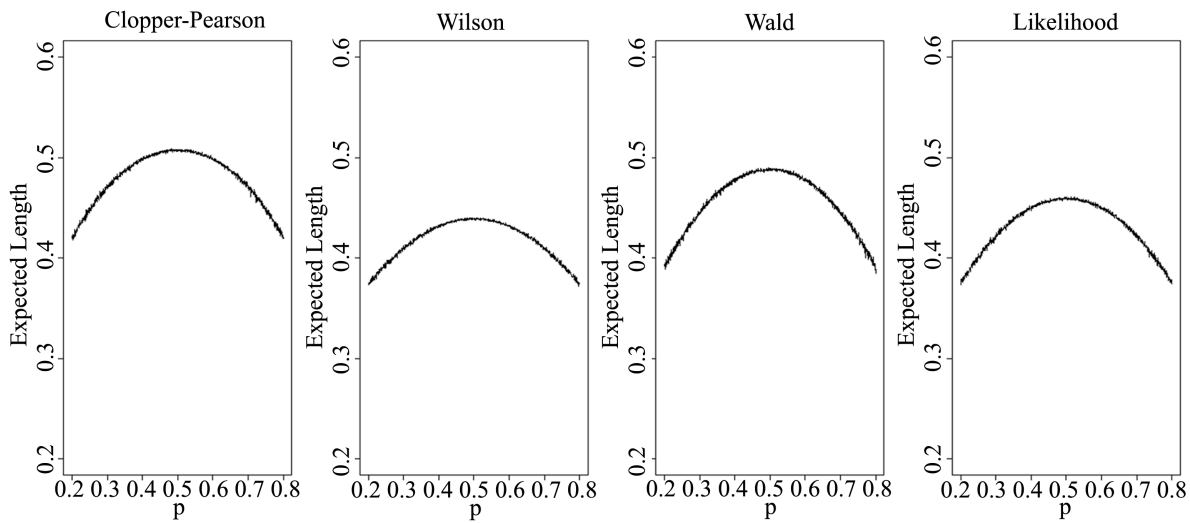
Figure 1. Coverage probabilities for $n = 15$.

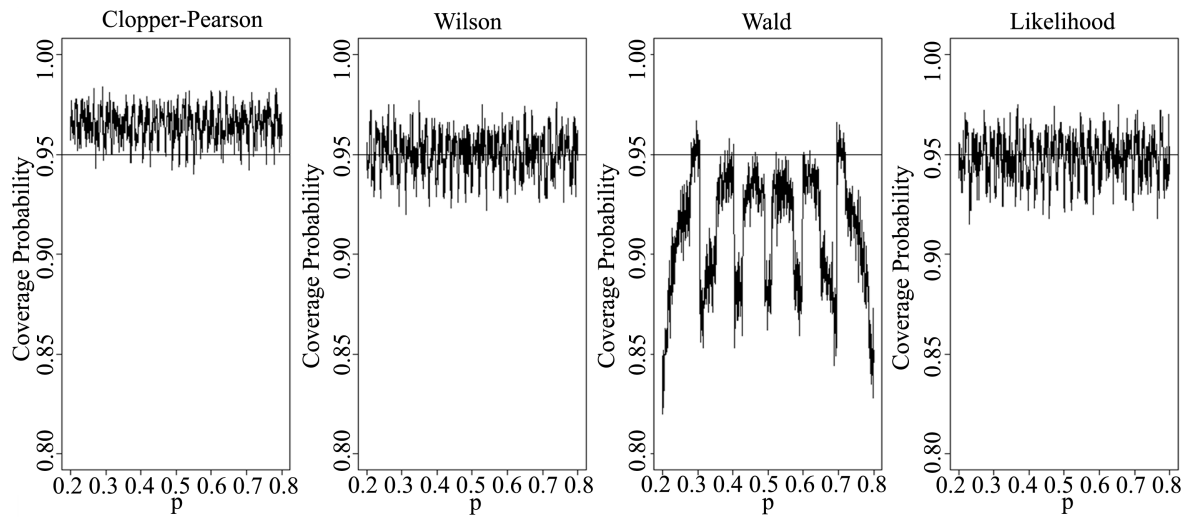**Figure 2.** Expected lengths for $n = 15$.

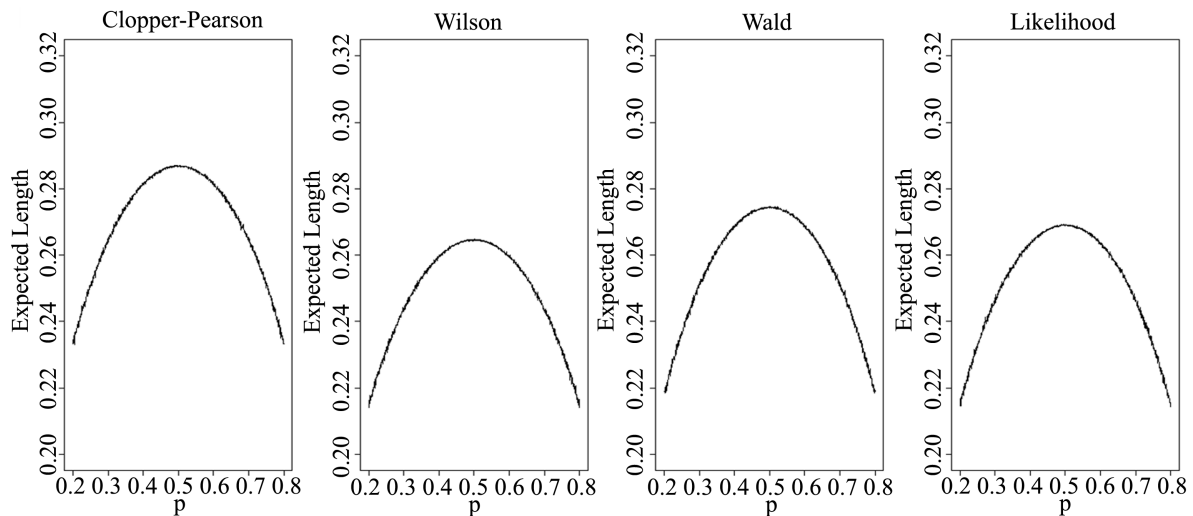**Figure 3.** Coverage probabilities for $n = 50$.

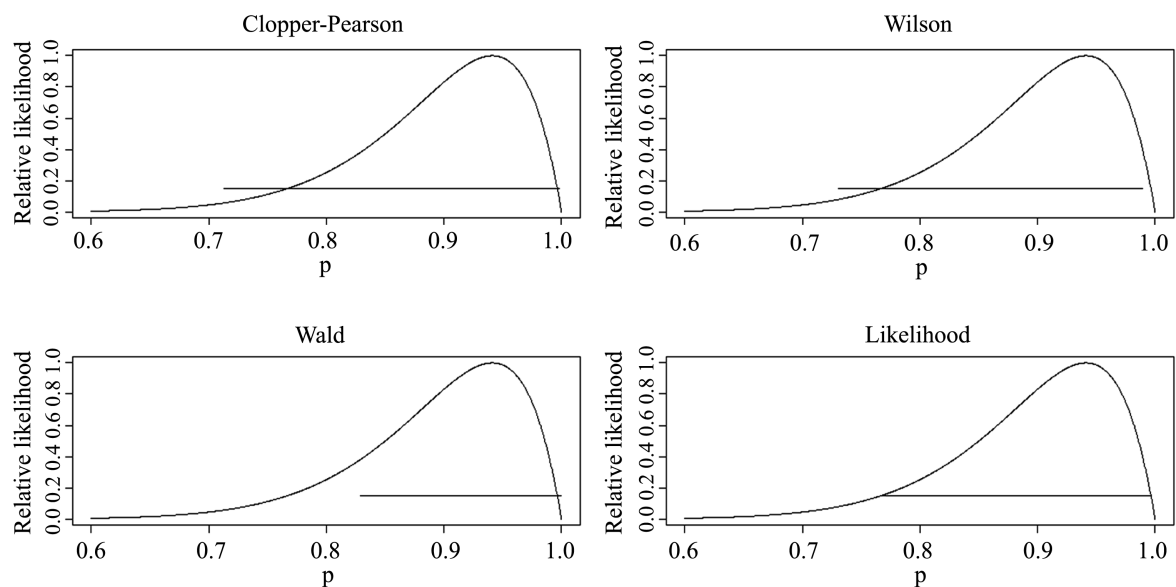**Figure 4.** Expected lengths for $n = 50$.

Figure 5. Confidence intervals plotted on the graph of relative likelihood function.

The likelihood interval looks optimal by evidence presented in Table 1. With these four confidence intervals we can conclude that hyperdynamic therapy is an effective method for treating ischaemic neurological symptoms due to vasospasm.

5. Conclusion

Clopper-Pearson interval is conservative for both small and large samples; however, it is always wider than it should. The Wald interval is well known and frequently used in statistical practice. Unfortunately, according to the above simulation study, its coverage probabilities are lower than the nominal level and are associated with problem of overshoot. Therefore, the inferential comparisons and judgements based on them might be misleading. On the other hand, Wilson

and Likelihood intervals have coverage probabilities near the nominal level and shorter lengths. Wilson interval for the real data application is wider than the likelihood interval and includes implausible values of the parameter. In summary, the Wilson and Likelihood intervals are recommended to be used in practice. It is worth noting the Likelihood interval looks superior to Wilson interval in that it is shorter and includes plausible values of the parameter p. The likelihood method has one drawback in the sense that it does not produce an interval when the number of successes x is 0 or n.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Matiri, G., Nyongesa, K. and Islam, A. (2017) Sequentially Selecting between Two Experiment for Optimal Estimation of a Trait with Misclassification. *American Journal of Theoretical and Applied Statistics*, **6**, 79-89. https://doi.org/10.11648/j.ajtas.20170602.12

[2] Vollset, S.E. (1993) Confidence Intervals for a Binomial Proportion. *Statistics in Medicine*, **12**, 809-824. https://doi.org/10.1002/sim.4780120902

[3] Agresti, A. and Coull, B.A. (1998) Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, **52**, 119-126. https://doi.org/10.1080/00031305.1998.10480550

[4] Newcombe, R.G. (1998) Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine*, **17**, 857-872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C857::AID-SIM777%3E3.0.CO;2-E

[5] Brown, L.D., Cai, T.T. and DasGupta, A. (2001) Interval Estimation for a Binomial Proportion. *Statistical Science*, **16**, 101-117. https://doi.org/10.1214/ss/1009213286

[6] Brown, L.D., Cai, T.T. and DasGupta, A. (2002) Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *Annals of Statistics*, **30**, 160-201. https://doi.org/10.1214/aos/1015362189

[7] Panagiotis, M. and Konstantinos, Z. (2008) Interval Estimation for a Binomial Proportion. A Bootstrap Approach. *Journal of Statistical Computation and Simulation*, **78**, 1251-1265. https://doi.org/10.1080/00949650701749356

[8] Kalbfleisch, J.G. (1985) Probability and Statistical Inference, Volume 2: Statistical Inference. 2nd Edition, Springer Verlag, New York. https://doi.org/10.1007/978-1-4612-5136-1

[9] Clopper, C.J. and Pearson, E.S. (1934) The Use of Confidence Intervals or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, **26**, 404-413. https://doi.org/10.1093/biomet/26.4.404

[10] Wilson, E.B. (1927) Probable Inference, the Law of Succession, and Statistical Inference. *Journal of American Statistical Association*, **22**, 209-212. https://doi.org/10.1080/01621459.1927.10502953

[11] Pritz, M.B., Zhou, X.H. and Brizendine, E.J. (1996) Hyperdynamic Therapy for Cerebral Vasospam: A Meta-Analysis of 14 Studies. *J. Neurovasc. Dis.*, **1**, 6-8.