OPTIMIZED CONFIDENCE INTERVALS FOR BINOMIAL DISTRIBUTED SAMPLES

SORANA D. BOLBOACĂ AND LORENTZ JÄNTSCHI

Abstract. The aim of the research was to develop an optimization procedure of computing confidence intervals for binomial distributed samples based. An inductive algorithm stands as method used to solve the problem of confidence intervals estimation for binomial proportions. The implemented optimization procedure uses two triangulations (varying simultaneously two pairs of three variables). The optimization method was assessed in a simulation study for a significance level of 5%, and sample sizes that vary from six to one thousand and associated possible proportions. The obtained results are available online [1]. Overall, the optimization method performed better, the values of cumulative error function decreasing in average with 10%, depending on the sample sizes and the confidence intervals method with which it is compared. The performances of the optimization method increase toghether with sample size, surprisingly because it is well known that the confidence interval methods that use the normal approximation hypothesis for a binomial distribution obtain good results with increasing of sample sizes.

1. Introduction

1.1. **History.** The origins of the mathematical study of natural phenomena are found in the fundamental work of ***Isaac Newton*** [1643-1727], *Philosophiae naturalis principia mathematica*, London, England, 1687. The mathematical basis of the binomial distribution study was put by ***Jacob Bernoulli*** [1654-1705], of which studies of especially significance for the theory of probabilities was published 8 years later after his death by his nephew, Nicolaus Bernoulli (*Ars Conjectandi*, Basel, Switzerland, 1713). In the *Doctrinam de Permutationibus & Combinationibus* section of this fundamental work he demonstrates the Newton binom. Later, ***Abraham De Moivre*** [1667-1754] put the basis of approximated calculus in the field, using the normal distribution for binomial distribution approximation (*The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play*, first edition in Latin published in Philosophical Transactions, by Royal Society, London, England, in 1711, second edition in English published by W. Pearforn, in 1738, which contain from

1991 *Mathematics Subject Classification.* 49M25, 60A05, 94B70, 62P10, 62Hxx.

Key words and phrases. Optimization, Confidence interval, Binomial distribution, Contingency table.

page 235 to page 243 the work entitled *Approximatio ad Summam Terminorum Binomii* $(a+b)^n$ *in Seriem expansi* presented privately to some friends in 1733). Later, with work *Theoria combinations observationum erroribus minnimis obnoxiae*, Comm. Soc. Reg. Scient., Got. Rec. Bd. V, IV, S. 1-53, 1823, ***Johann Carl Friedrich Gauss*** [1777-1855] put the basis of mathematical statistics. ***Abraham Wald*** [1902-1950, born in Cluj] do his contributions on confidence intervals study, elaborated and published the confidence interval that has his name now, in the paper *Contributions to the Theory of Statistical Estimation and Testing Hypothesis*, The Annals of Mathematical Statistics, p. 299-326, 1939. Nowadays, the most prolific researcher on confidence intervals domain is ***Allan Agresti***, which it was named the *Statistician of the Year* for 2003 by *American Statistical Association*, and at the prize ceremony (October 14, 2003) it spoken about *Binomial Confidence Intervals*.

1.2. **Today.** As which as was related, the binomial distribution has origins in nature phenomena studies. A series of papers can be cited demonstrative in this sense. Thus, law of binomial distribution are proved at heterometric bands of tetrameric enzyme in [2], the stoichiometry of the donor and acceptor chromophores implied in enzymatic ligand/receptor interactions in [3], translocation and exfoliation of type I restriction endonucleases in [4], biotinidase activity on neonatal thyroid hormone stimulator in [5], the parasite induced mortality at fish in [6], the occupancy/activity for proteins at multiple nonspecific sites containing replication in [7]. Let note here the paper [8], which defines very well the frame and limits of binomial distribution model applied to the natural phenomena.

1.3. **Aim.** The formal definition of a confidence interval is *a confidence interval gives an estimated range of values that is likely to include an unknown population parameter, the estimated range being calculates from a given set of sample data*. If independent sample are taken repeatedly from population same, and a confidence interval for each sample is calculated, then a certain percentage (confidence level) of the intervals will include the unknown population parameter. Confidence intervals are usually computed for 95% confidence level. In experimental sciences, usually the scientist use a sample of size N from the entire population to test its hypothesis. Thus, the scientist it operate with a discrete variable X, which can collect its property of interest from the entire sample of size N. The statistical hypothesis for this variable is that are binomial distributed. Confidence interval estimations (CIE) for proportions using normal approximation has commonly uses for a simple fact: the normal approximation is easier to use in practice comparing with other approximate estimators [9]. Expressing of the true confidence intervals can be a metter of dead or alive. Let's say that in medicine with one percent a pacient can be killed.

The aim of this research: to obtain the binomial confidence intervals optimized boundaries for N less or equal with 1000, based on the original method of double

triangulation, and to assess the performances of these optimized confidence intervals compared with the well known exact methods.

2. Material and Method

2.1. **Background.** The problem of the CIE for a binomial proportion and binomial sample sizes is presented in many papers [10, 11]. It is well known that for low proportions, the lower confidence boundaries is frequently less than zero while for the proportions closer to the upper boundaries exceed one [12, 13]. The main problem of the existent methods is represented by the inadequate coverage and inappropriate intervals [13]. A series of previous papers of authors cummulates the knowledge from the literature relating to the usage and assessment of the binomial confidence intervals. From this series, of interest to the mathematical model used for this research are: [14], [15], and [16].

2.2. **Optimization Algorithm.** An inductive algorithm is proposed method used to solve the problems of CIE for binomial proportions. The implemented optimization procedure use two triangulations (vary simultaneously two pairs of three variables). For a given sample size (N), the program uses 34 starting points and optimization pathways in the optimization obtained from 17 different methods and variants of direct calculation of confidence interval limits, selected from literature. The optimization procedure makes changes at one or more unknowns (from 1 to 6) from n if the pathway change produces decreasing of cumulative error function value. On assessment of 95% confidence intervals for sample sizes varying from 7 to 1000 and all possible proportions the proposed optimization method was compared with the exact formulas.

2.3. **Assessment Procedure.** Twelve assessment methods were defined, extending five previously reported [16]. The formulas are below, where M is the method of CIE, N is the sample size, and ε^M is observed experimental error using method M, α is the imposed error level (usualy 5%) and all assessment methods depend on both M and N, and let's take A a binary variable (0 or 1):

$$AvgOEA = \sum_{X=A}^{N-A} \frac{\varepsilon_{X,N}^{M}}{N+1-2A},\ StDOEA = \sqrt{\sum_{X=A}^{N-A} \frac{(\varepsilon_{X,N}^{M} - AvgOEA)^2}{N-2A}}$$

$$SiDOEA = \sqrt{\sum_{X=A}^{N-A} \frac{(\varepsilon_{X,N}^{M} - 100\alpha)^2}{N+1-2A}},\ AvADAA = \sum_{X=A}^{N-A} \frac{|\varepsilon_{X,N}^{M} - AvgOEA|}{N-2A}$$

$$AvADSA = \sum_{X=A}^{N-A} \frac{|\varepsilon_{X,N}^{M} - 100\alpha|}{N+1-2A},\ S8DOEA = \sqrt[8]{\sum_{X=A}^{N-A} \frac{(\varepsilon_{X,N}^{M} - 100\alpha)^8}{N-2A}}$$

The assessment methods are constructed as follows: AvgOE0 from AvgOEA for A = 0; AvgOE1 from AvgOEA for A = 1; similarly for the remaining ones.

TABLE 1. Assessment of CI methods for $\alpha = 5\%$

N=20							
Method	Wald	A_C	Wils	Logit	Jeff	OptB	Bnt
AvgOE0	12.0	3.40	4.25	3.81	4.67	4.95	5
AvgOE1	10.8	3.08	3.85	3.45	4.22	4.48	5
StDOE0	9.35	1.02	1.67	1.22	1.49	1.38	0
StDOE1	9.57	1.41	2.03	1.63	1.99	1.98	0
SiDOE0	11.5	1.88	1.78	1.68	1.49	1.34	0
SiDOE1	11.0	2.36	2.29	2.22	2.10	2.00	0
AvADA0	6.84	0.93	1.28	0.96	1.24	1.24	0
AvADA1	6.84	1.22	1.54	1.31	1.53	1.49	0
AvADS0	7.06	1.60	1.60	1.41	1.23	1.18	0
AvADS1	6.86	1.92	1.93	1.76	1.59	1.54	0
S8DOE0	23.5	2.67	2.30	2.61	2.58	1.94	0
S8DOE1	23.2	3.76	3.74	3.75	3.75	3.73	0
N=200							
AvgOE0	6.28	4.73	5.00	4.90	5.07	5.07	5
AvgOE1	6.22	4.68	4.95	4.85	5.02	5.02	5
StDOE0	3.65	0.86	0.75	0.78	0.77	0.60	0
StDOE1	3.69	0.98	0.89	0.91	0.92	0.78	0
SiDOE0	3.86	0.90	0.74	0.78	0.77	0.61	0
SiDOE1	3.88	1.02	0.89	0.92	0.92	0.78	0
AvADA0	1.49	0.65	0.59	0.59	0.48	0.42	0
AvADA1	1.49	0.70	0.63	0.63	0.53	0.46	0
AvADS0	1.34	0.63	0.58	0.58	0.48	0.42	0
AvADS1	1.37	0.68	0.63	0.63	0.53	0.46	0
S8DOE0	17.9	2.11	1.70	1.92	2.09	1.69	0
S8DOE1	17.9	2.84	2.82	2.83	2.84	2.82	0

Wald: Wald; A_C: Agresti-Coull; Wils: Wilson;
Logit: Logit; Jeff: Jeffreys; OptB: Optimized;
Bnt: Best is near to.

3. Results and Discussion

Table 1 contains the assessment results for two values of sample size N. The most important observation is that OptB is strictly monotone on S8DOE0, which was in fact (after its discovering) the objective function in optimization. Another interesting remark is based on statistical significance of the result. Thus, coming back to our population of size N is clearly that if we extract X subjects from it, and if we want to express a confidence interval for this selection, we cannot expect to something like 2.3, because we cannot extract 2.3 subjects from the sample!

Thus, based on this remark, the subject of the optimization was only natural numbers from 0..N interval, when "magic numbers" were obtained for N=20 and α=5% (Table 2, Ξ).

Table 3 gives the experimental error obtained with OptB for every X for N=20, and following formulas were used to convert obtained "magic numbers" to confidence interval boundaries (N=20 in Table 3), where *eps.* meaning a

TABLE 2. OptB "magic numbers" for N=20 and α =5%

i	0	1	2	3	4	5	6	7	8	9
Ξ_i	0	0	1	1	2	2	3	4	5	5
i	10	11	12	13	14	15	16	17	18	19
Ξ_i	6	7	8	9	10	11	12	14	15	17

TABLE 3. Experimental error of OptB for all X for N=20

X	0,20	1,19	2,18	3,17	4,16	5,15	6,14	7,13	8,12	9,11	10
$\epsilon_{\cdot,20}^{OptB}$	0.00	7.55	4.32	6.07	4.37	6.52	5.26	3.17	3.70	4.03	4.14

The error is symetrically distributed;
above are pairs of X excepting the center (10).

small subunitary value (i.e. an open boundary, see Table 2):

$$CI_{Lower}(0,N)=0;\ CI_{Upper}(N,N)=N;$$

$$CI_{Lower}(N-i+1)=\Xi_{N-i}+eps.,\quad i=\overline{1,N}$$

$$CI_{Upper}(i-1)=N-\Xi_{N-i+1},\quad i=\overline{1,N}$$

4. Conclusions

Optimized confidence intervals for simple proportion (X from N) in binomial distribution hypothesis were obtained and assessed. A strictly monotone asssessment method (S8DOE0) was discovered and later used for obtaining of all boundaries of confidence intervals for N varying from 2 to 1000, results being available online.

5. Further Plans

Based on the methodology developed for simple proportion, optimized confidence intervals for other types of formulas (including here all formulas which are used as medical key parameters) computed on 2×2 contingency tables are subject to further investigation of authors.

6. Acknowledgments

This research was partly supported by UEFISCSU Romania through project ET46/2006 and CNCSIS Romania through project AT93/2007.

References

[1] Jäntschi L, Bolboaca SD. http://l.academicdirect.org/Statistics/binomial_distribution, ©2005, June

[2] Engel W. Onset of synthesis of mitochondrial enzymes during mouse development. Synchronous activation of parental alleles at the gene locus for the M form of NADP dependent malate dehydrogenase, HUMANGENETIK 1973 20(2):133-140

[3] Meadows DL, Schultz JS. A molecular model for singlet/singlet energy transfer of monovalent ligand/receptor interactions. Biotechnology and Bioengineering 1991 37(11):1066-1075

[4] Szczelkun MD. Kinetic models of translocation, head-on collision, and DNA cleavage by type I restriction endonucleases. Biochemistry 2002 41(6):2067-2074

[5] Tanyalcin T, Eyskens F, Philips E, Lefevere M, Buyukgebiz B. A marked difference between two populations under mass screening of neonatal TSH and biotinidase activity. Accreditation and Quality Assurance 2002 7(11):498-506

[6] Osset EA, Fernandez M, Raga JA, Kostadinov A, Mediterranean Diplodus annularis (Teleostei: Sparidae) and its brain parasite: Unforeseen outcome. Parasitology International 2005 54(3):201-206

[7] Conant CR, Van Gilst MR, Weitzel SE, Rees WA, von Hippel PH, A Quantitative Description of the Binding States and In Vitro Function of Antitermination Protein N of Bacteriophage? Journal of Molecular Biology 2005 348(5):1039-1057

[8] Carlton MA, Stansfield WD. Making babies by the flip of a coin? American Statistician, 2005 59(2):180-182

[9] Pawlikowski DC, McNickle GE, Coverage of Confidence Intervals in Sequential Steady-State Simulation. Simul Pract Theor, 1998 6:255-267

[10] Borkowf CB. Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. Stat Med 2006 25(21):3679-3695

[11] Reiczigel J. Confidence intervals for the binomial parameter: some new considerations. Stat Med 2003 22(4):611-621

[12] Newcombe RG. Two-sided confidence intervals for the single proportion; comparison of several methods. Stat Med 1998 17:857-872

[13] Brown DL, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Stat Sci 2001 16:101-133

[14] Drugan T, Bolboaca SD, Jäntschi L, Achimas Cadariu BA. Binomial Distribution Sample Confidence Intervals Estimation 1. Sampling and Medical Key Parameters Calculation. Leonardo Electronic Journal of Practices and Technologies 2003 2(3):45-74

[15] Bolboaca SD, Achimas Cadariu BA. Binomial Distribution Sample Confidence Intervals Estimation 2. Proportion-like Medical Key Parameters, Leonardo Electronic Journal of Practices and Technologies 2003 2(3):75-110

[16] Bolboaca SD, Jäntschi L. Binomial Distribution Sample Confidence Intervals Estimation for Positive and Negative Likelihood Ratio Medical Key Parameters. Annual Symposium on Biomedical and Health Informatics. American Informatics Medical Association, Special Issue: from Foundations, to Applications to Policy, Bethseda MD USA 2005:66-70

"Iuliu Hatieganu" University of Medicine and Pharmacy, 400349 Cluj, Romania
E-mail address: sorana@j.academicdirect.ro

Technical University of Cluj-Napoca, 400641 Cluj, Romania
E-mail address: lori@academicdirect.org