Are confidence intervals for binomial distributed samples an optimization meters?

SORANA-DANIELA BOLBOACĂ[1], LORENTZ JÄNTSCHI[2]

[1] "Iuliu Hațieganu" University of Medicine and Pharmacy
13 Emil Isac Street, 400023 Cluj-Napoca, Romania
http://sorana.cademicdirect.ro
sorana@j.academicdirect.ro

[2] Technical University of Cluj-Napoca,
15 Constantin Daicoviciu Street, 400020 Cluj-Napoca, Romania
http://lori.academicdirect.org
lori@j.academicdirect.org

Key Words and Phrases: Optimization, Confidence interval, Binomial distribution, Contingency table

AMS Subject Classification: 49M25 (Discrete approximations), 60A05 (Axioms; other general questions), 94B70 (Error probability), 62P10 (Applications to biology and medical sciences), 62Hxx (Multivariate analysis)

The aim of the research was to develop an optimization procedure of computing confidence intervals for binomial distributed samples based.

An inductive algorithm is proposed 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 at the following address:

http://l.academicdirect.org/Statistics/binomial_distribution/

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 with increasing of the 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.

Are confidence intervals for binomial distributed samples an optimization meters?

Sorana-Daniela BOLBOACĂ

and

Lorentz JÄNTSCHI

History

- Origins
  - ***Newton’s binom (a+b)^N***
  - fundamental work of *Isaac NEWTON* [1643-1727], *Philosophiae naturalis principia mathematica*, London, England, 1687
  - ***Bernoulli distribution and binomial distribution***
  - The mathematical basis of the binomial distribution study was put by Jacob BERNOULLI [1654-1705]; studies were 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

History

- ***Binomial distribution approximating***
- *Abraham DeMOIVRE* [1667-1754] use the normal distribution for binomial distribution approximation: *The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play*, 1-st edition in Latin (Philosophical Transactions, Royal Society, London, England, 1711), 2-nd edition in English (W. Pearforn, 1738) - contain from 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
- ***Gauss distribution***
- The work *Theoria combinations observationum erroribus minnimis obnoxiae*, Comm. Soc. Reg. Scient., Got. Rec. Bd. V, IV, S. 1-53, 1823 of *Johann Carl Friedrich GAUSS* [1777-1855]

History

- ***Wald confidence interval***
- *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: *Contributions to theTheory of Statistical Estimation and Testing Hypothesis*, The Annals of Mathematical Statistics, p. 299-326, 1939
- ***Agresti-Coull confidence interval***
- *Nowadays, the most prolific researcher on confidence intervals study 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.*

Origins and relevance

- Binomial distribution has origins in nature phenomena studies
- Demonstrative in this sense, binomial distribution are proved at heterometric bands of tetrameric enzyme [1], the stoichiometry of the donor and acceptor chromophores implied in enzymatic ligand/receptor interactions [2], translo-cation and exfoliation of type I restriction endonucleases [3], biotinidase activity on neonatal thyroid hormone stimulator [4], the parasite induced mortality at fish [5], the occupancy/activity for proteins at multiple non-specific sites containing replication [6]
- The paper [7], defines very well the frame and limits of binomial distribution model applied to the natural phenomena

A formal definition

- *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, and a confidence interval (CI) for each sample is calculated, then a certain percentage (confidence level, CL) of the intervals will include the unknown population parameter.
- CI are usually computed for 95% CL.

Motivation

- In experimental sciences, usually the scientist use a sample of given size 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 [8].
- Expressing of the true confidence intervals can be a matter of dead or alive. Let’s say that in medicine with one percent a patient can be killed.

Our aim

- 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
  - to assess the performances of these optimized confidence intervals compared with the well known methods

Background

- For the problem of the CIE for a binomial proportion and binomial sample sizes see papers [9, 10].
- Problems using exact formula methods [11, 12]:
  - For low proportions: lower confidence boundary - frequently less than zero;
  - for high proportions: upper boundaries – frequenty exceed one;
  - Main problem of existent methods - inadequate coverage and inappropriate intervals [12].
- Our results relating to the usage and assessment of the binomial confidence intervals: [13], [14], and [15].

Optimization algorithm

- An inductive algorithm were developed to solve the CIE for binomial proportions.
- The 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.

Assessment procedure

- Twelve assessment methods were defined, extending five previously reported [15].
- The notations for the formulas given in the next:
  - M - method of CIE
  - N - sample size
  - ε^M - observed experimental error using method M
  - α - the imposed error level (usually 5%)
- All assessment methods depend on both M and N
- A is a binary variable (0 or 1)

Assessment formulas

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

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

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

Setting assessment formulas

- Given formulas has A letter at the end of the name; A take two values: 0 and 1
- We have 12 assessment formulas:
  - AvgOE0 (*α*), AvgOE1 (*α*)
  - StDOE0 (0), StDOE1 (0)
  - SiDOE0 (0), SiDOE1 (0)
  - AvADA0 (0), AvADA1 (0)
  - AvADS0 (0), AvADS1 (0)
  - S8DOE0 (0), S8DOE1 (0)
- In the parenthesis are given the “best of” for every formula

CI boundaries and OptB “magic numbers”

- For a 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!
- Confidence boundaries for *X* from *N must be natural numbers less or equal to N*.
- Based on this observation, algorithm were implemented. Next table gives these so called magic numbers (Ξ) for *N = 20* and *α = 5%*, the next are formulas and the last table are with observed errors for *N = 20* and *α = 5%* where note that X=1 and X=19 has same observed error due to complementarity's, and same for the rest of.

CI boundaries and OptB “magic numbers”

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

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

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

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

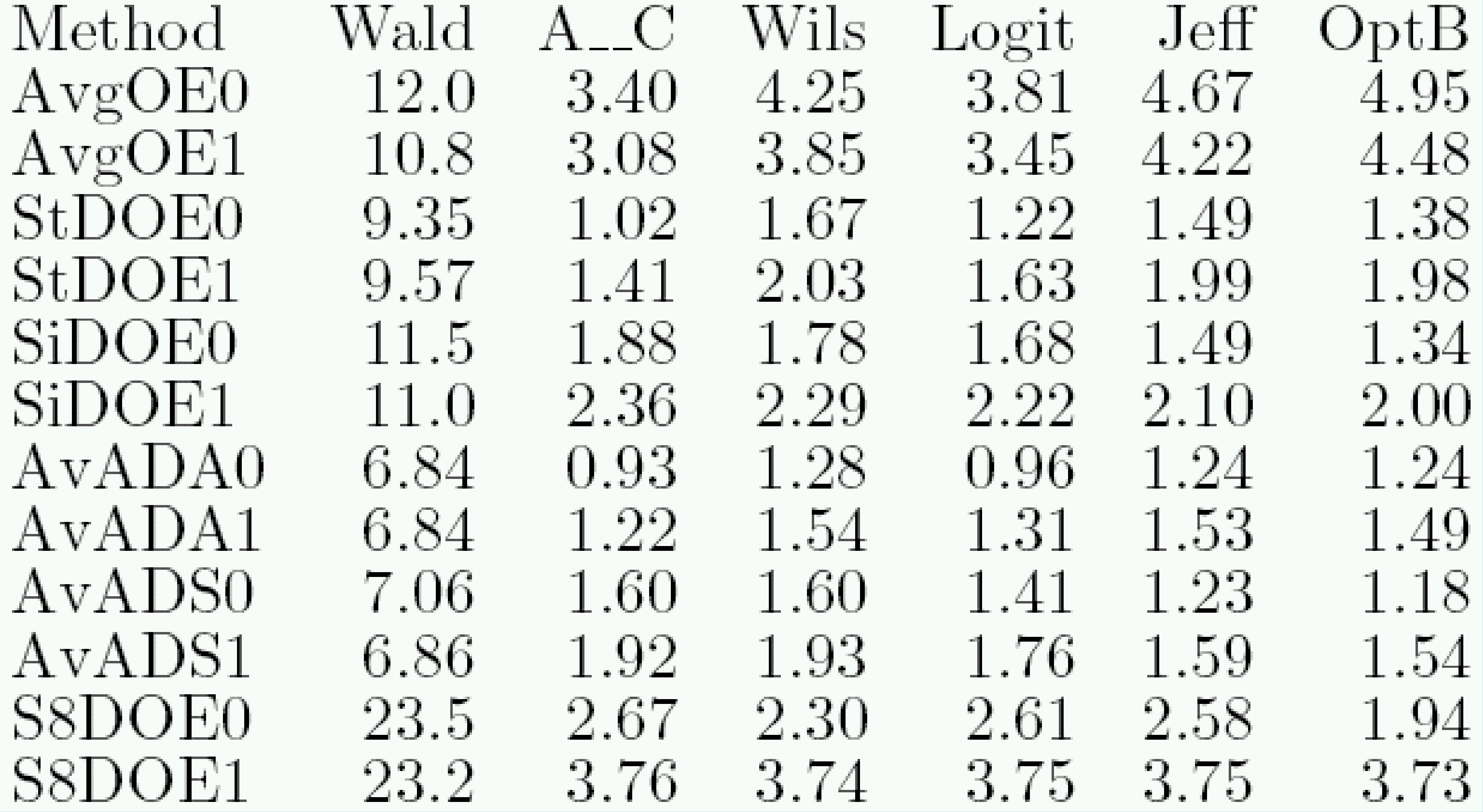
X	0,20	1,19	2,18	3,17	4,16	5,15	6,14	7,13	8,12	9,11	10
ε	0.00	7.55	4.32	6.07	4.37	6.52	5.26	3.17	3.70	4.03	4.14

Results - algorithm

OptB	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
X=0	100	**0**	**0**	**0**	**0**	**0**	**0**	**0**
X=1	33.99	39.66	19.83	**5.51**	**0.92**	**0.09**	**0.01**	**0**
X=2	9.49	26.56	31.87	21.25	8.5	**2.04**	**0.27**	**0.02**
X=3	**1.99**	10.44	23.5	29.38	22.03	9.91	**2.48**	**0.27**
X=4	**0.27**	**2.48**	9.91	22.03	29.38	23.5	10.44	**1.99**
X=5	**0.02**	**0.27**	**2.04**	8.5	21.25	31.87	26.56	9.49
X=6	**0**	**0.01**	**0.09**	**0.92**	**5.51**	19.83	39.66	33.99
X=7	**0**	**0**	**0**	**0**	**0**	**0**	**0**	100

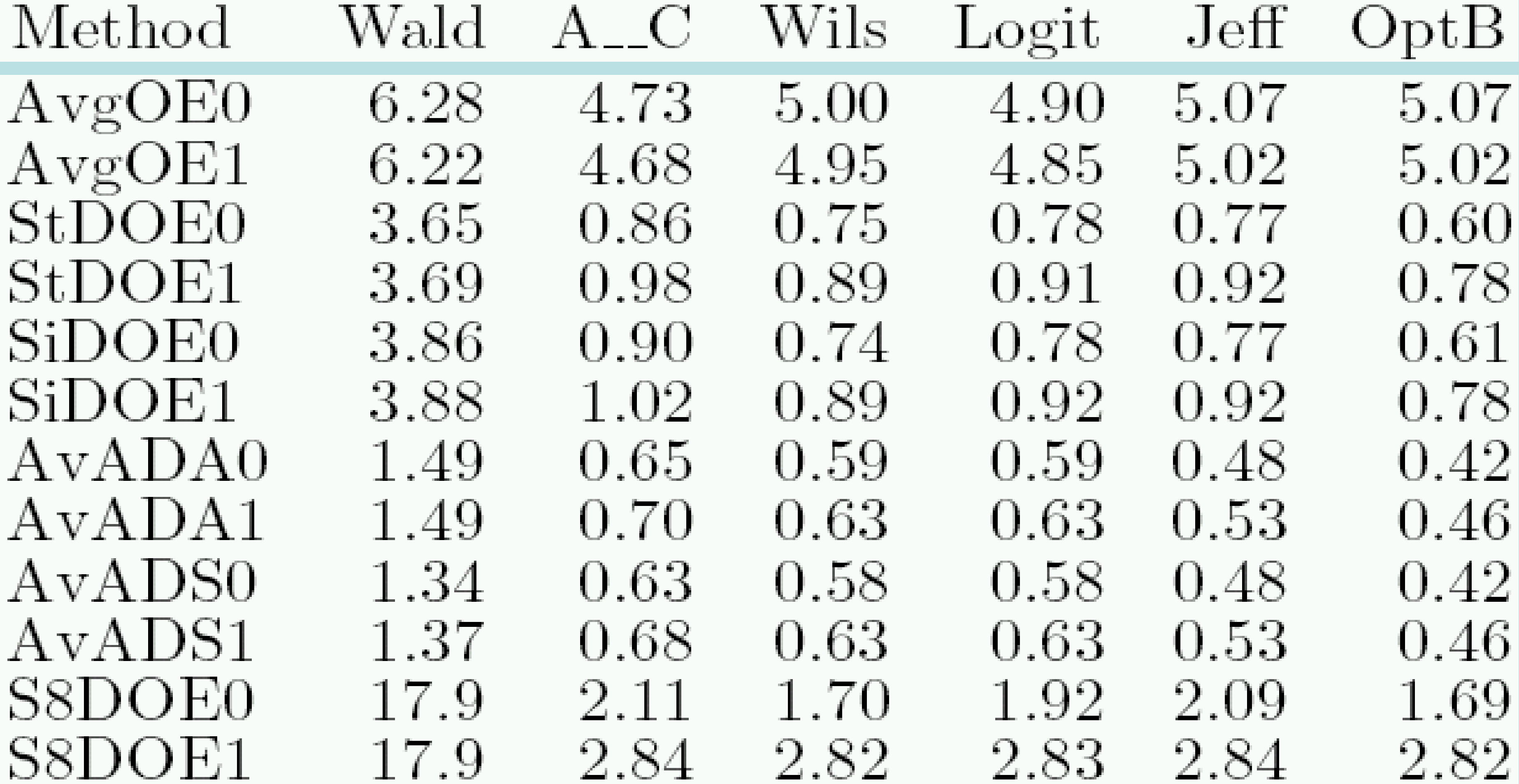
The table gives the selecting of confidence level boundaries by the optimization procedure (OptB) for N=7 and $\alpha = 5\%$.

Assessment of CI for N=20, α=5%



Method	Wald	A__C	Wils	Logit	Jeff	OptB
AvgOE0	12.0	3.40	4.25	3.81	4.67	4.95
AvgOE1	10.8	3.08	3.85	3.45	4.22	4.48
StDOE0	9.35	1.02	1.67	1.22	1.49	1.38
StDOE1	9.57	1.41	2.03	1.63	1.99	1.98
SiDOE0	11.5	1.88	1.78	1.68	1.49	1.34
SiDOE1	11.0	2.36	2.29	2.22	2.10	2.00
AvADA0	6.84	0.93	1.28	0.96	1.24	1.24
AvADA1	6.84	1.22	1.54	1.31	1.53	1.49
AvADS0	7.06	1.60	1.60	1.41	1.23	1.18
AvADS1	6.86	1.92	1.93	1.76	1.59	1.54
S8DOE0	23.5	2.67	2.30	2.61	2.58	1.94
S8DOE1	23.2	3.76	3.74	3.75	3.75	3.73

Assessment of CI for N=200, α=5%



Method	Wald	A__C	Wils	Logit	Jeff	OptB
AvgOE0	6.28	4.73	5.00	4.90	5.07	5.07
AvgOE1	6.22	4.68	4.95	4.85	5.02	5.02
StDOE0	3.65	0.86	0.75	0.78	0.77	0.60
StDOE1	3.69	0.98	0.89	0.91	0.92	0.78
SiDOE0	3.86	0.90	0.74	0.78	0.77	0.61
SiDOE1	3.88	1.02	0.89	0.92	0.92	0.78
AvADA0	1.49	0.65	0.59	0.59	0.48	0.42
AvADA1	1.49	0.70	0.63	0.63	0.53	0.46
AvADS0	1.34	0.63	0.58	0.58	0.48	0.42
AvADS1	1.37	0.68	0.63	0.63	0.53	0.46
S8DOE0	17.9	2.11	1.70	1.92	2.09	1.69
S8DOE1	17.9	2.84	2.82	2.83	2.84	2.82

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:

http://l.academicdirect.org/

└ Statistics/binomial_distribution/

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.

Acknowledgments

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

- Thank you for your attention!

References (1/3)

[1] 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

[2] 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

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

[4] 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

[5] 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

References (2/3)

[6] 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

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

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

[9] 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

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

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

References (3/3)

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