Optimal Data Analysis
Vol. 3 (October 10, 2014), 110-112

2155-0182/10/$3.00

Discrete 95% Confidence Intervals for ODA Model- and Chance-Based Classifications

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

This article describes and illustrates discrete 95% confidence intervals (CIs) which are computed in novometric analysis for both model- and chance-based classification results.

In the parametric general linear model and maximum likelihood statistical paradigms, 95% CIs are conceptualized as continuous random variables distributed in the manner dictated by the assumed underlying parent population.[1,2] In contrast to this parametric conceptualization of the nature of classical data, the fundamental premise of both the optimal ("maximum accuracy") data analysis (ODA) paradigm[3,4] and novometric theory[5,6] is that classical phenomena are fundamentally discrete in nature, and no reference is made concerning a hypothetical underlying parent distribution. After briefly discussing conceptual similarity between central tenants of novometry and quantum mechanics (QM), the discrete constitution of CIs used in novometric analysis is illustrated using an empirical example.

Novometry, QM, and Information Theory

It is important to understand conceptual consistencies, and also to distinguish between novometric theory and QM, a field of physics that derives from the finding that some physical phenomena change in discrete amounts (Latin: *quanta*), rather than along a continuum.[7] In QM information is measured by von Neumann entropy, a generalization of classical information theory (a mathematical model of communication) applied to quantum phenomena.[8] A widely employed unit of quantum information is a binary system known as a *qubit*, and the information content of a message is measured in terms of the minimum number of binary models (*n* qubits) required to store the message. This is analogous to a *bit* in binary-log classical information theory, where the mean number of bits needed to store or communicate one symbol in a message is called entropy.

In QM the *Particle in a Box* model of energy held in a confined space is conceptually analogous to the premise of sample strata in novometry. According to this QM model the energy of a particle in infinite space has continuous solutions, but as constraints are imposed, such as physical confinement, discrete solutions occur which represent the only possible solutions. For single confined particle systems these solutions represent discrete energy levels.

Copyright 2014 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

In novometric theory the size of the box corresponds to the amount of data (N) in the sample (the population corresponds to the universe), and the discrete measurement levels correspond to sample strata: patient strata for classifying disease incidence; market segments for classifying consumer preference; or storm categories for classifying drilling platform structural damage, for example.[3-4] The set of all possible solutions (which in novometry is called the *descendant family*) for a given application (sample) is identified in novometry using the minimum denominator selection search algorithm (MDSA).[5,6] Also conceptually reminiscent of information theory, novometry involves the use of binary *parses* to create sample strata, and the information content of a model relating two variables (a class variable is modeled using one or more attributes) is measured in terms of the minimum number of strata (model endpoints) required to achieve the best combination of accuracy and parsimony in the classification of the phenomenon.[3-6]

Discrete CIs in Novometry

Discrete 95% CIs in novometric analysis for model- and chance-based classification are illustrated using recent research investigating the relationship between the temperature of an Emergency Department (ED) patient and the disease status [community-acquired pneumonia (CAP) or influenza-like illness (ILI)] of the patient.[9] MDSA identified a descendant family of two possible solutions for the sample of 200 patients which is presented in Table 1.

Bootstrap methodology using 10,000 iterations of a 50% resample with replacement[3] is used to obtain the discrete 95% CI for model classification accuracy. For this demonstration the analysis was performed on the two-strata model. The resulting estimates of ESS were sorted and cumulated, yielding the results presented in Table 2. As seen in Table 2, the 5% bound of the discrete 95% CI for the model is rounded as 26.0 (highlighted in yellow), and the 95% bound (highlighted in yellow) is rounded as 56.5 (see also Table 1).

Table 1: Summary of MDSA Procedure for Discriminating CAP and ILI Patients

Step	Strata	MinD	ESS	Efficiency
1	4	37	46.9 30.9-62.4 0.82-15.4	11.7 7.72-15.6 0.21-3.84
2	2	88	41.4 26.0-56.5 0.19-14.2	20.7 13.0-28.2 0.10-7.12

Note: There were two steps in this MDSA.[9] Strata is the number of partitions identified by the CTA model. MinD is the smallest number of observations (patients) in any of the strata (i.e., the smallest model endpoint N). ESS is a normed index of classification accuracy on which 0 represents the level of accuracy expected by chance and 100 represents perfect (errorless) classification. By rule-of-thumb: ESS<25 is a relatively weak effect; ESS<50 is a moderate effect; ESS<75 is a relatively strong effect; and ESS>75 is a very strong effect.[3] Efficiency, an index of parsimony, is ESS/number of strata. Under the ESS and Efficiency point estimates, the first row is the exact discrete 95% CI for the model, and the second row is the corresponding 95% CI for chance. Highlight is used to indicate how tabled values were obtained: yellow for the model effect (see also Table 2), and blue for the chance effect (see also Table 3).

Table 2: Cumulated Results of Model Bootstrap

Quantile	Estimate
100% Max	74.85
99%	62.60
95%	56.48
90%	53.25
75% Q3	48.99
50% Median	41.67
25% Q1	35.50
10%	29.58
5%	25.96
1%	19.65
0% Min	7.39

Copyright 2014 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

For the chance effect 10,000 iterations of Fisher’s randomization procedure is performed using a 50% resample with replacement, and the results are cumulated (see Table 3).

Table 3: Cumulated Results of Chance Fisher’s Randomization

Quantile	Estimate
100% Max	31.13
99%	18.76
95%	14.25
90%	12.19
75% Q3	8.06
50% Median	4.32
25% Q1	2.25
10%	0.19
5%	0.19
1%	0.19
0% Min	0.19

As seen in Table 3, the 5% bound of the discrete 95% CI for the model (highlighted in blue) is given as 0.19 (see Table 1), and the 95% bound (highlighted in blue) is rounded as 14.2 (see Table 1).

References

[1] Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, DC, APA Books.

[2] Grimm LG, Yarnold, P.R. (2000). *Reading and understanding more multivariate statistics*. Washington, DC, APA Books.

[3] Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

[4] Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

[5] Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

[6] Yarnold PR, Soltysik RC (2014). Globally optimal statistical models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3*, 78-84.

[7] Shanker R (1994). *Principles of quantum mechanics, 2nd edition*. New York, NY, Springer.

[8] Wilde MM (2013). *Quantum information theory*. Cambridge, UK, Cambridge University Press.

[9] Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, *3*, 102-106.

Author Notes

Author contributions were described earlier.[4]

E-mail: Journal@OptimalDataAnalysis.com.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646