Optimal Data Analysis
Vol. 2, Release 2 (November 27, 2013), 220-221

2155-0182/10/$3.00

MegaODA Large Sample and BIG DATA Time Trials: Maximum Velocity Analysis

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

This third time trail of newly-released MegaODA™ software studies the fastest-to-analyze application, known as a 2x2 cross-classification table. Designs involving unweighted binary data are arguably currently the most widely employed across quantitative scientific disciplines as well as engineering fields including communications, graphics, data compression, real-time processing and autonomous synthetic decision-making, among others. The present simulation research is run on a 3 GHz Intel Pentium D microcomputer and reveals MegaODA returns the exact one- or two-tailed Type I error rate, as well as all of the other classification-relevant statistics provided in UniODA analysis, in fractions of a CPU second for samples of a million observations.

The first time trial of MegaODA software[1] used Monte Carlo (MC) simulation to ascertain that the software is capable of rapidly ruling-out effects that are *not* statistically significant (*ns*) for applications involving ordered attributes, or for categorical attributes having five response categories. Rule-in analysis of ordered attributes was assessed in a second time trial[2], but hasn't yet been reported for categorical attributes. Accordingly the present study presents initial time trials for rule-out and rule-in analysis conducted for the most common categorical design, the 2x2 cross-classification table. In this study the special-purpose TABLE algorithm which is available in UniODA and MegaODA software, designed to maximize solution speed, is used to conduct the analyses, and MC simulation is not run because exact *p* is computed (ODA software randomization and Fisher's exact test are isomorphic for binary applications).[3]

Table 1: Experimental Data for Study of Categorical Attributes: Large Sample

	Weak Effect	
	Attribute	
Class Variable	0	1
0	25,000	25,000
1	20,000	30,000
	Moderate Effect	
	Attribute	
Class Variable	0	1
0	35,000	15,000
1	15,000	35,000

Seen in Table 1, the experimental study of categorical data involved four 2x2 tables featuring *weak* (ESS=10.0, ESP=10.1) and

2155-0182/10/$3.00

moderate (ESS=ESP= 40.0) effects[3] for designs with n=100,000: multiply each tabled value by ten to get the data used for n=10^6. The following UniODA and MegaODA code[3] was used to analyze these 2x2 tables (program commands are indicated in red).

```
OPEN DATA;
OUTPUT weak100K.out;
CATEGORICAL ON;
TABLE 2;
CLASS ROW;
DATA;
25000 25000
20000 30000
END DATA;
GO;
```

All analyses involving large or BIG DATA samples, and weak or moderate effects, were completed in fractions of a CPU second: the use of 0 CPU seconds reported in software output indicates that less than half of a CPU second elapsed. Chaining 100 sequential runs of the largest problem and timing overall solution time via stopwatch suggested that each analysis required approximately 0.15 CPU seconds to initiate, run and complete. Development of a special-purpose "pipeline" architecture and the use of other supercomputing methods[4] would further enhance problem solution speed by more than an order of magnitude, should the need for such speed ever arise.

References

[1]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, 2, 194-197.

[2]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the wheat. *Optimal Data Analysis*, 2, 202-205.

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

[4]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences, 22*, 739-752.

Author Notes

ODA Blog: odajournal.com