Optimal Data Analysis
Vol. 2, Release 2 (November 19, 2013), 194-197

2155-0182/10/$3.00

MegaODA Large Sample and BIG DATA Time Trials: Separating the Chaff

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

Just-released MegaODA™ software is capable of conducting UniODA analysis for an unlimited number of attributes using samples as large as one million observations. To minimize the computational burden associated with Monte Carlo simulation used to estimate the Type I error rate (*p*), the first step in statistical analysis is identifying effects that are *not* statistically significant or *ns*. This article presents an experimental simulation exploring the ability of MegaODA to identify *ns* effects in a host of designs involving a binary class variable, under ultimately challenging discrimination conditions (all data are random) for sample sizes of n=100,000 and n=1,000,000. Most analyses were solved in CPU seconds running MegaODA on a 3 GHz Intel Pentium D microcomputer. Using MegaODA it is straightforward to rapidly rule-out *ns* effects using Monte Carlo simulation with BIG DATA for large numbers of attributes in simple or complex, single- or multiple-sample designs involving categorical or ordered attributes either with or without weights being applied to individual observations.

For this research a data set was constructed with three independently generated random numbers provided for each of 10^6 observations. The first variable was binary, created using a random probability value (*p*) generated from a uniform distribution: BINARY=0 if $p \leq 0.05$; BINARY=1 if $p > 0.05$. The second variable was an ordered 5-point Likert-type scale[1] created using another random *p* generated from a uniform distribution: LIKERT=1 if $p < 0.2$; LIKERT=2 if $0.2 \leq p < 0.4$; LIKERT=3 if $0.4 \leq p < 0.6$; LIKERT=4 if $0.6 \leq p < 0.8$; and LIKERT=5 if $p > 0.8$. The third and final variable is a constrained real-number created as yet another random probability value generated from a uniform distribution: RANDOM=*p*.

Observations were assigned a unique ID number, an integer between 1 and 10^6 used to segment the sample: two analyses are reported, the first for observations with ID≤100,000, and the second for the complete sample. The trials begin using the former sample, which might be described as being relatively large as compared with the typical empirical research published in peer-reviewed literature.

Relatively Large Samples

For the sample of n=100,000, BINARY was evenly distributed with 49,978 Class 0 and 50,022 Class 1 observations. Descriptive statis-

2155-0182/10/$3.00

tics for LIKERT and RANDOM data are given in Table 1 separately by BINARY, the class variable in analyses presented in this research.

Table 1: Descriptive Statistics for LIKERT and RANDOM Data by Class: n=100,000

Variable	Statistic	Class 0	Class 1
LIKERT	Mean	2.996	3.012
	SD	1.418	1.412
	Median	3	3
	Skewness	0.001	-0.014
	Kurtosis	-1.308	-1.297
RANDOM	Mean	0.501	0.500
	SD	0.288	0.290
	Median	0.501	0.497
	Skewness	-0.003	0.007
	Kurtosis	-1.197	-1.210

Categorical attribute. In the first pair of time trials LIKERT was treated as a *categorical attribute* with five response categories. Widely seen across quantitative fields this configuration is called a rectangular categorical design.[2]

The *post hoc* hypothesis that class can be discriminated using *categorical* LIKERT scores was tested by running the following MegaODA code (control commands are indicated in red):

```
OPEN total.dat;
OUTPUT total.out;
VARS id binary likert random;
CLASS binary;
ATTR likert;
CAT likert;
EX id>100000;
MC ITER 1000 TARGET .05 STOPUP 99.9;
GO;
```

Monte Carlo simulation is parameterized to target generalized "per-comparison" p<0.05 for a single test of a statistical hypothesis.[1] A maximum allowance of 1000 MC experiments is indicated. STOPUP ceases simulation when specified confidence that estimated p exceeds target p is achieved.[1] Analysis stopped after 100 MC iterations: estimated p<0.23, confidence for target p>0.10 is >99.9%. With negligible ESS (0.65) and ESP (0.68), the model required <1 CPU second to solve (analyses herein were run on a 3 GHz Intel Pentium D microcomputer). A LOO "leave-one-out" jackknife analysis[1] that was conducted in a separate run finished in less than one CPU second.

The *post hoc* hypothesis that class can be discriminated using *categorical* LIKERT scores, with observations *weighted* by a *continuous* variable (RANDOM), was tested by adding the following code.

```
WEIGHT random;
GO;
```

Analysis stopped at 200 MC iterations: estimated p<0.12, confidence for target p>0.05 is >99.99%. Having a negligible weighted ESS (0.91) and ESP (0.95), the model needed 1 CPU second to finish. LOO analysis is not available for weighted categorical designs.[1]

Ordinal attribute. The second set of two time trials treated LIKERT as being an *ordinal* attribute with five discrete response categories, another design that is prevalently used across quantitative disciplines.

The *post hoc* hypothesis that class can be discriminated via *ordinal* LIKERT scores was tested by using the prior MegaODA code with WEIGHT and CAT commands commented-out.

```
*CAT likert;
*WEIGHT random;
GO;
```

Analysis stopped in 300 MC iterations: estimated p<0.094; confidence for target p>0.05 is >99.9%. With negligible ESS (0.65) and ESP (0.68), the model took 2 CPU seconds to finish (the model identified was identical to the model for the corresponding categorical analysis).

2155-0182/10/$3.00

The *post hoc* hypothesis that class can be discriminated with *ordinal* LIKERT scores, with observations *weighted* by a *continuous* variable (RANDOM), was tested by amending the code.

```
WEIGHT random;
GO;
```

Analysis stopped at 1000 MC iterations: estimated $p<0.026$, confidence for target $p>0.05$ is >0.1%. With negligible weighted ESS (0.91) and ESP (0.95), the model needed 6 CPU seconds to complete. MC simulation also reported confidence for target $p<0.05$ is >99.99%.

Continuous attribute. The third set of two time trials treated RANDOM as being a *continuous* attribute, another design frequently used across quantitative disciplines.

The *post hoc* hypothesis that class can be discriminated via *continuous* RANDOM scores was tested by prior code amended as shown.

```
ATTR random;
*WEIGHT random;
GO;
```

Analysis stopped in 100 MC iterations: estimated $p<0.21$; confidence for target $p>0.10$ is >99.9%. With negligible ESS (0.65) and ESP (0.79) the model took 99 CPU seconds to finish.

The *post hoc* hypothesis that class can be discriminated using *continuous* LIKERT scores, with observations *weighted* by an *ordinal* variable (RANDOM), amending the code.

```
WEIGHT likert;
GO;
```

Analysis stopped in 100 MC iterations: estimated $p<0.72$, confidence for target $p>0.10$ is >99.99%. Having a negligible weighted ESS (0.66) and ESP (0.80), the model needed 49 CPU seconds to finish.

Multiple-sample analysis. The final time trial for large samples used the UniODA Generalizability (Gen) procedure which identifies the UniODA model that—when it is simultaneously applied to multiple samples, maximizes the *minimum ESS* achieved by the model across classes.[1] The *post hoc* hypothesis that class can be discriminated by a single model involving a *continuous* attribute (RANDOM) independently applied to five different *samples* (LIKERT) was tested via the following MegaODA code.

```
CLASS binary;
ATTR random;
GEN likert;
GO;
```

Analysis stopped in 100 MC iterations: estimated $p<0.33$; confidence for target $p>0.10$ is >99.99%. Having negligible ESS (0.49) and ESP (1.13) the model required 78 CPU seconds.

BIG DATA Samples

For the sample of $n=10^6$, BINARY was evenly distributed: 499,928 Class 0 and 500,072 Class 1 observations. Descriptive statistics for LIKERT and RANDOM data are given in Table 2 separately by BINARY, the class variable in analyses presented below.

Table 2: Descriptive Statistics for LIKERT and RANDOM Data by Class: $n=10^6$

Variable	Statistic	Class 0	Class 1
LIKERT	Mean	2.997	3.003
	SD	1.414	1.415
	Median	3	3
	Skewness	0.001	-0.004
	Kurtosis	-1.300	-1.301
RANDOM	Mean	0.500	0.500
	SD	0.289	0.289
	Median	0.501	0.500
	Skewness	-0.002	0.001
	Kurtosis	-1.200	-1.202

2155-0182/10/$3.00

All seven time trial experiments run for the large sample are also run for the BIG DATA sample. MegaODA code is the same as used previously, with EX commented-out.

```
*EX ID>100000;
```

Categorical attribute. Analysis for the *post hoc* hypothesis that class can be discriminated using *categorical* LIKERT scores stopped after 100 MC iterations: estimated $p<0.19$, confidence for target $p>0.05$ is >99.99%. With negligible ESS (0.20) and ESP (0.21), the model required 3 CPU seconds to solve. LOO analysis conducted in a separate run required less than 1 CPU second to complete.

Analysis for the *post hoc* hypothesis that class can be discriminated using *categorical* LIKERT scores, with observations *weighted* by a *continuous* variable (RANDOM) stopped after 200 MC iterations: estimated $p<0.12$, confidence for target $p>0.05$ is >99.99%. With negligible weighted ESS (0.26) and ESP (0.27), the model required 9 CPU seconds to solve.

Ordinal attribute. Analysis for the *post hoc* hypothesis that class can be discriminated using *ordinal* LIKERT scores stopped after 300 MC iterations: estimated $p<0.107$, confidence for target $p>0.05$ is >99.99%. With negligible ESS (0.20) and ESP (0.21), the model required 42 CPU seconds to solve.

Analysis for the *post hoc* hypothesis that class can be discriminated using *ordinal* LIKERT scores, with observations *weighted* by a *continuous* variable (RANDOM) stopped after 1000 MC iterations: estimated $p<0.049$, confidence for target $p>0.05$ is >47.69%, confidence for $p<0.05$ is >52.31%. Having a negligible weighted ESS (0.26) and ESP (0.27), the model required 144 CPU seconds to solve. This is an example of the worst-case (most resource intensive) scenario. At this point the attribute would be subjected to a much more computationally intensive MC simulation[3] as will be studied in forthcoming rule-in time trials.

Continuous attribute. Analysis for the *post hoc* hypothesis that class can be discriminated via *continuous* RANDOM scores stopped after 14 MC iterations: estimated $p<0.58$, confidence for target $p>0.05$ is >99.99%. With negligible ESS (0.16) and ESP (0.18), the model required 140 CPU seconds to solve.

Analysis for the *post hoc* hypothesis that class can be discriminated using *continuous* RANDOM scores, with observations *weighted* by an *ordinal* variable (LIKERT) stopped after 28 MC iterations: estimated $p<0.86$, confidence for target $p>0.10$ is >99.99%. With negligible weighted ESS (0.17) and ESP (0.18), the model required 127 CPU seconds to solve.

Multiple-sample analysis. Analysis for the *post hoc* hypothesis that class can be discriminated using a single model involving a *continuous* attribute (RANDOM) which is independently applied to five different *samples* (LIKERT) stopped after 12 MC iterations: estimated $p<0.17$, confidence for target $p>0.05$ is >99.98%. With negligible ESS (0.09) and ESP (0.11), the model required 131 CPU seconds.

References

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

[2]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190.

[3]Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo estimation of two-category UniODA two-tailed p. Optimal Data Analysis, 1, 43-45.

Author Notes

ODA Blog: http://odajournal.com