Optimal Data Analysis
Vol. 2, Release 2 (November 11, 2013), 177-190

2155-0182/10/$3.00

Univariate and Multivariate Analysis of Categorical Attributes with Many Response Categories

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

A scant few weeks ago disentanglement of effects identified in purely categorical designs in which all variables are categorical, including notoriously-complex rectangular categorical designs (RCDs) in which variables have a different number of response categories, was poorly understood. However, univariate and multivariate optimal ("maximum-accuracy") statistical methods, specifically UniODA and automated CTA, make the analyses of such designs straightforward. These methods are illustrated using an example involving *n*=1,568 randomly selected patients having either confirmed or presumed *Pneumocystis carinii* pneumonia[1] (PCP). Four categorical variables used in analysis include patient status (two categories: alive, dead), gender (male, female), city of residence (seven categories), and type of health insurance (ten categories). Examination of the cross-tabulations of these variables makes it obvious why conventional statistical methods such as chi-square analysis, logistic regression analysis, and log-linear analysis are both inappropriate for, as well as easily overwhelmed by such designs. In contrast, UniODA and CTA identified maximum-accuracy solutions effortlessly in this application.

Univariate Analysis: Efficient UniODA-Based *Range Tests* Boost Statistical Power

Data for a random sample of 1,568 PCP patients are *gender* (male=1, female=2), *status* (alive=0, dead=1), *city* of residence (Los Angeles or LA=1, Chicago=2, New York or NY=3, Seattle=4, Miami=5, Nashville=6, Phoenix=7); and type of insurance or *insure* (1=Medicaid, 2=Medicare, 3 is an unused code, 4=fee for service, 5=PPO, 6=POS, 7=managed care, 8=HMO, 9=private non-HMO, 10=self-pay, 11=charitable organization). Univariate results are obtained for all six of the bivariate pairings of these four variables.

Status and *Gender*. Table 1 presents the 2x2 cross-tabulation of status and gender.

Table 1: Status and Gender

	Females	Males
Alive	278	937
Deceased	23	100

2155-0182/10/$3.00

The *post hoc* hypothesis that women and men had a different mortality rate was tested by running the following UniODA[2] code (control commands are indicated using red):

```
VARS gender status city insure;
CLASS status;
ATTR gender;
CATEGORICAL gender;
GO;
```

Monte Carlo simulation was not used to estimate Type I error in this analysis because in non-weighted 2x2 binary designs the UniODA randomization algorithm and Fisher's exact test are isomorphic.[2] The resulting model was: if gender=female predict status=alive; if gender= male predict status=dead. The model was not statistically significant (exact $p>0.10$) and it had negligible accuracy (ESS=4.2) and predictive value (ESP=2.0).[3] There is no evidence that men and women had different mortality rates.

Status and *City*. Table 2 presents the 2x7 cross-tabulation of status and city.

Table 2: Status and City

	Alive	Deceased
LA	190	12
Chicago	250	15
NY	316	44
Seattle	165	17
Miami	133	16
Nashville	95	10
Phoenix	66	9

The *post hoc* hypothesis that cities had different mortality rates was tested by running the following appended UniODA code:

```
ATTR city;
CAT city;
MC ITER 10000;
GO;
```

The resulting model was: if city=LA or Chicago predict status=alive; for all other cities predict status=dead. The model was statistically significant (estimated $p<0.035$, confidence for $p<0.05$ is >99.99%), with weak accuracy (ESS= 14.3) and negligible predictive value (ESP=5.2). Table 3 presents the resulting confusion table.

Table 3: Confusion Table for UniODA Model Predicting Status Based on City

		Predicted Status		
		Alive	Deceased	
Actual	Alive	440	775	36.2%
Status	Deceased	27	96	78.1%
		94.2%	11.0%	

Findings thus far may be symbolically indicated as: (LA, Chicago) > Rest; where parentheses indicate it hasn't yet been determined if embedded cities are significantly different on status; > indicates a significantly greater proportion of living patients; and Rest indicates all other cities in the sample.

The second step of this UniODA *range-test* procedure involves two comparisons, one between LA and Chicago, and another between the other five cities: Monte Carlo simulation is thus parameterized to target experimentwise $p<0.05$ using the Sidak criterion for three tests of statistical hypotheses (two forthcoming tests and the initial test[2]). UniODA code for the first test is appended as follows:

```
EX city>2;
MC ITER 10000 TARGET .05 SIDAK 3;
GO;
```

The resulting model was not statistically significant (exact $p>0.10$), with minute accuracy (ESS=1.3) and predictive value (ESP=0.3), and so the symbolic representation of the effect thus far remains unchanged.

2155-0182/10/$3.00

The UniODA code for the second test is replaced as follows:

```
EX city<3;
GO;
```

The resulting model was not statistically significant (confidence for $p>0.10$ is >99.99%), with negligible accuracy (ESS=5.9) and minute predictive value (ESP=2.3), and so the symbolic representation of the effect is complete. There is evidence that the mortality rate is comparable in LA and Chicago, and is significantly lower than the (comparable) mortality rates in NY, Seattle, Miami, Nashville, and Phoenix.

To conduct all-possible comparisons for pairs of these seven cities would require running and integrating (7*6)/2=21 analyses, with final runs using a SIDAK criterion for 21 tests. The UniODA range-test procedure used 3 tests. The SIDAK criterion for 3 versus 21 tests is target $p<0.017$ and $p<0.0025$, respectively.[2]

Status and *Insurance*. Table 4 gives the 2x10 cross-tabulation of status and insurance.

Table 4: Status and Insurance

	Alive	Deceased
Medicaid	127	16
Medicare	68	6
Fee for Service	78	4
PPO	92	8
POS	471	49
Managed Care	32	5
HMO	92	10
Private non-HMO	52	5
Self-Pay	38	5
Charitable Group	165	15

The *post hoc* hypothesis that different types of insurance had different mortality rates was tested running the following UniODA code:

```
ATTR insure;
CAT insure;
MC ITER 1000;
GO;
```

The resulting UniODA model was: if insurance=Medicare, fee for service, PPO, private non-HMO, or charitable group then predict status=alive; for all other insurance categories predict status=dead. The model was not statistically significant (confidence for $p>0.10$ is >99.99%), with negligible accuracy (ESS=6.6) and predictive value (ESP=2.4). There thus is no evidence that different types of insurance are associated with different mortality rates.

To conduct all-possible comparisons for pairs of these ten insurance categories requires running and integrating (10*9)/2=45 analyses, with final runs using a SIDAK criterion for 45 tests, $p<0.00114$. The UniODA range-test procedure used one test, with target $p<0.05$.

Gender and *City*. Table 5 gives the 2x7 cross-tabulation of gender and city.

Table 5: Gender and City

	Females	Males
LA	18	184
Chicago	41	224
NY	116	244
Seattle	78	104
Miami	11	138
Nashville	25	80
Phoenix	12	63

The *post hoc* hypothesis that women and men had different mortality rates in different cities was tested using this UniODA code:

```
CLASS gender;
ATTR city;
CAT city;
MC ITER 10000;
GO;
```

The resulting UniODA model was: if city=LA, Chicago, Miami, or Phoenix then predict gender=male; for all other cities predict

2155-0182/10/$3.00

gender=female. The model was statistically significant (estimated $p<0.0001$, confidence for $p<0.05$ is >99.99%), with moderate accuracy (ESS =31.5) and weak predictive value (ESP=22.0). Table 6 is the resulting confusion table.

Table 6: Confusion Table for UniODA Model Predicting Gender Based on City

		Predicted Gender		
		Male	Female	
Actual	Male	609	428	58.7%
Gender	Female	82	219	72.8%
		88.1%	33.8%	

Findings thus far are symbolically indicated with respect to proportion of females as: (NY, Seattle, Nashville)>(LA, Chicago, Miami, Phoenix).

The second step of this UniODA *range-test* procedure involves two comparisons, one between New York, Seattle, and Nashville, and another between LA, Chicago, Miami and Phoenix: Monte Carlo simulation is thus parameterized to target experimentwise $p<0.05$ using the Sidak criterion for three tests of statistical hypotheses (two forthcoming tests and the initial test). UniODA code for the first test is appended as follows:

```
EX city=3;
EX city=4;
EX city=6;
GO;
```

The resulting UniODA model was: if city=LA or Miami predict gender=male; if city= Chicago or Phoenix predict gender=female. The model was statistically significant (estimated $p<0.0081$, confidence for experimentwise $p<0.05$ is >99.99%), with weak accuracy (ESS= 17.5) and negligible predictive value (ESP=7.3). Table 7 gives the resulting confusion table.

Table 7: Confusion Table for UniODA Model, City and Gender: LA, Chicago, Miami, Phoenix

		Predicted Gender		
		Male	Female	
Actual	Male	322	287	52.9%
Gender	Female	29	53	64.6%
		91.7%	15.6%	

The symbolic representation of the effect thus far is: (NY, Seattle, Nashville) > (Chicago, Phoenix) > (LA, Miami). UniODA code for the second test is replaced as follows:

```
EX city<3;EX city=5;EX city=7;GO;
```

The resulting UniODA model was: if city=NY or Nashville then predict gender=male; if city=Seattle then predict gender=female. The model was statistically significant (confidence for experimentwise $p<0.05$ is >99.99%), with weak accuracy (ESS=11.3) and predictive value (ESP=12.5). Table 8 gives the confusion table.

Table 8: Confusion Table for UniODA Model, City and Gender: NY, Seattle, Nashville

		Predicted Gender		
		Male	Female	
Actual	Male	324	104	75.7%
Gender	Female	141	78	35.6%
		69.7%	42.9%	

The symbolic representation of the effect thus far is: Seattle > (NY, Nashville) > (Chicago, Phoenix) > (LA, Miami).

The third step of this UniODA *range-test* procedure involves three comparisons, one for comparison for each set of parentheses remaining in the symbolic representation. Monte

2155-0182/10/$3.00

Carlo simulation is thus parameterized to target experimentwise $p<0.05$ using the Sidak criterion for six tests of statistical hypotheses (the three forthcoming tests and the prior three tests). UniODA code for the first test is replaced:

```
EX city=1;EX city=2;EX city=4;
EX city=5;EX city=7;GO;
```

The resulting model was not statistically significant (confidence for $p>0.10$ is >99.99%), with negligible accuracy (ESS=7.0) and predictive value (ESP=8.4), and so the symbolic representation thus far remains unchanged. Similar results were obtained for the second and third tests so the symbolic representation is complete.

Obtaining the confusion table for this final UniODA model requires integrating confusion tables for the two halves of the analysis. By adding corresponding entries in confusion tables for the first (Table 7) and second (Table 8) analyses, the integrated table is created, as is seen in Table 9: overall ESS=5.8 and ESP=4.3.

Table 9: Confusion Table for Final UniODA Model Predicting Gender Based on City

		Predicted Gender		
		Male	Female	
Actual	Male	646	391	62.3%
Gender	Female	170	131	43.5%
		79.2%	25.1%	

There is evidence that the proportion of females in the sample is significantly greater in Seattle than in NY or Nashville (which are statistically comparable), which have a significantly greater proportion of female patients in the sample than Chicago or Phoenix (and are statistically comparable), which have a significantly greater proportion of female patients than LA or Miami.

To conduct all-possible comparisons for pairs of seven cities requires running and integrating 21 analyses, with final runs using a SIDAK criterion of target $p<0.00244$. In contrast the UniODA range-test procedure used six tests for target $p<0.008513$.

Gender and *Insurance*. Table 10 is the 2x7 cross-tabulation of gender and insurance.

Table 10: Gender and Insurance

	Females	Males
Medicaid	18	125
Medicare	17	57
Fee for Service	12	70
PPO	8	92
POS	152	368
Managed Care	8	29
HMO	25	77
Private non-HMO	12	45
Self-Pay	14	29
Charitable Group	35	145

The *post hoc* hypothesis that women and men had different types of insurance coverage was tested using the following UniODA code:

```
CLASS gender;
ATTR insure;
CAT insure;
MC ITER 1000;
GO;
```

The resulting UniODA model was: if insurance=Medicaid, fee for service, PPO, managed care, private non-HMO or charitable group then predict gender=male; if insurance=Medicare, POS, HMO, or self-pay then predict gender=female. The model was statistically significant (estimated $p<0.0001$, confidence for $p<0.01$ is >99.99%), with weak accuracy (ESS=17.9) and predictive value (ESP=12.6). Table 11 presents the resulting confusion table.

Findings thus far are symbolically indicated with respect to proportion of females as:

2155-0182/10/$3.00

(Medicare, POS, HMO, self-pay)>(Medicaid, fee for service, PPO, managed care, private non-HMO, charitable group).

Table 11: Confusion Table for UniODA Model Predicting Gender Based on Insurance

		Predicted Gender		
		Male	Female	
Actual	Male	506	531	48.8%
Gender	Female	93	208	69.1%
		84.5%	28.2%	

The second step of this UniODA range-test procedure involves two comparisons, one comparison for each set of parentheses: Monte Carlo simulation is thus parameterized to target experimentwise $p<0.05$ using the Sidak criterion for three tests of statistical hypotheses (two forthcoming tests and the initial test). UniODA code for the first test is appended as follows:

```
EX insure=1;EX insure=4;EX insure=5;
EX insure=7;EX insure=9;EX insure=11;
MC ITER 1000 TARGET .05 SIDAK 3;
GO;
```

The resulting model was not statistically significant (confidence for $p>0.10$ is >99.99%), with negligible accuracy (ESS=5.0) and predictive value (ESP=5.6), thus symbolic representation remains unchanged. UniODA code for the second test is appended as follows:

```
EX insure=2;EX insure=6;EX insure=8;
EX insure=10;GO;
```

The resulting UniODA model was: if insurance=Medicaid, fee for service, or PPO then predict gender=male; if insurance=managed care, private non-HMO, or charitable group predict gender=female. The model was not statistically significant at the experimentwise criterion, but it met the generalized "per-comparison" criterion for $p<0.05$ (confidence >99.99%): ESS=15.9, ESP=8.4. The symbolic notation is thus complete, unless it is decided to include the effect significant at the generalized criterion, in which case final symbolic notation would be: (Medicare, POS, HMO, self-pay)>(Medicaid, fee for service, PPO)>(managed care, private non-HMO, charitable group). There is evidence that females in the sample are *most* (comparably) likely to have Medicare, POS, HMO, or self-pay health coverage; significantly (comparably) *less* likely to have Medicaid, fee for service, or PPO health coverage; and significantly (comparably) *least* likely to have managed care, private non-HMO, or charitable group health coverage. The UniODA range-test tested three statistical hypotheses, versus 45 needed for all possible comparisons.

City and *Insurance*. The final univariate analysis, Table 12 is the 7x10 cross-tabulation of city and insurance. Cell entries indicated in red are very small and render analysis by chi-square, logistic regression analysis, log-linear model, and other maximum-likelihood-based methods unsuitable because the minimum expectation is too small in too many cells.[4-6]

Not presented, the UniODA model (with CLASS city; ATTR insure;) was statistically significant using 1000 Monte Carlo experiments (estimated $p<0.001$, confidence>99.99% for target $p<0.01$), and had moderate accuracy (ESS= 39.0) and predictive value (ESP=41.5). However, no observations were predicted by the model to reside in Seattle.

Table 12: City and Insurance

	LA	Chi	NY	Sea	Mia	Nas	Pho
Mcaid	43	23	61	5	8	0	3
Mcare	23	15	5	19	8	0	4
FFS	36	16	1	7	15	2	5
PPO	11	26	29	7	19	1	7
POS	68	64	257	73	58	0	0
MCare	0	0	0	0	0	0	37
HMO	0	0	0	0	0	102	0
nHMO	5	19	2	20	0	0	11
Self	7	19	3	5	4	0	5
Charity	9	83	2	46	37	0	3

2155-0182/10/$3.00

No algorithmic procedure has yet been developed to disentangle effects found in such *supercategorical designs* involving two or more categorical attributes each with response scales consisting of three or more categories. Based on the present analysis, there nevertheless is evidence that type of health insurance coverage is not comparably represented across cities.

Multivariate Analysis: CTA is Optimal, Logistic Regression Analysis Overwhelmed

Exposition turns to multivariate analyses in purely categorical designs, and two analyses are planned. In the first analysis status will be treated as the class variable and predicted using gender, city and insurance as possible attributes, and in the second analysis gender will be treated as the class variable and predicted using status, city and insurance as possible attributes.

Compared with analytically troublesome data seen in Table 12, cross-tabulation results in Table 13 might well be described as "the end of the linear statistical analysis world." (Non)linear classification methods from the general linear model and the maximum-likelihood paradigms maximize variance ratios or the value of the likelihood function for the sample, respectively.[5,6] A problem presented by the present data for these methods is satisfying the multivariate normally distributed (MND) assumption required for *p* to be valid. As for Table 12, cell entries indicated in red are very small and render analysis by chi-square, logistic regression analysis, log-linear model, and other maximum-likelihood-based methods unsuitable because the minimum expectation is too small in too many cells.[4-6] Some logistic regression analysis software systems add the value 0.5 to every cell in an effort to circumvent division by zero due to affected matrices being less than full rank.[5,6] If done presently this would require adding the equivalent of 156*0.5 or 78 observations to the sample: 5.0% of the actual total *n*.

Table 13: Distribution of Four Cross-Tabulated Categorical Variables Investigated Presently

GENDER	STATUS	CITY	INSURANCE	*n*
Female	Alive	Los Angeles	Medicaid	0
			Medicare	2
			Fee for Service	2
			PPO	0
			POS	8
			Managed Care	0
			HMO	0
			Private non-HMO	2
			Self Pay	1
			Local Charity	1
		Chicago	Medicaid	0
			Medicare	3
			Fee for Service	2
			PPO	1
			POS	14
			Managed Care	0
			HMO	0
			Private non-HMO	1
			Self Pay	7
			Local Charity	12
		New York	Medicaid	15
			Medicare	3
			Fee for Service	1
			PPO	4
			POS	79
			Managed Care	0
			HMO	0
			Private non-HMO	1
			Self Pay	1
			Local Charity	1
		Seattle	Medicaid	0
			Medicare	8
			Fee for Service	3
			PPO	2
			POS	33
			Managed Care	0
			HMO	0
			Private non-HMO	5
			Self Pay	2
			Local Charity	19
		Miami	Medicaid	1
			Medicare	0
			Fee for Service	2
			PPO	1
			POS	5
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	1
			Local Charity	0
		Nashville	Medicaid	0
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	0
			Managed Care	23
			HMO	0
			Private non-HMO	0
			Self Pay	0

2155-0182/10/$3.00

			Local Charity	0
		Phoenix	Medicaid	0
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	0
			Managed Care	8
			HMO	0
			Private non-HMO	1
			Self Pay	1
			Local Charity	1
	Deceased	Los Angeles	Medicaid	0
			Medicare	0
			Fee for Service	1
			PPO	0
			POS	1
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	0
			Local Charity	1
		Chicago	Medicaid	0
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	0
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	0
			Local Charity	0
		New York	Medicaid	1
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	9
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	1
			Local Charity	0
		Seattle	Medicaid	0
			Medicare	1
			Fee for Service	0
			PPO	0
			POS	3
			Managed Care	0
			HMO	0
			Private non-HMO	2
			Self Pay	0
			Local Charity	0
		Miami	Medicaid	1
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	0
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	0
			Local Charity	0
		Nashville	Medicaid	0
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	0
			Managed Care	0
			HMO	2
			Private non-HMO	0
			Self Pay	0
			Local Charity	0
		Phoenix	Medicaid	0
			Medicare	0
			Fee for Service	0
			PPO	0
			POS	0
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	0
			Local Charity	0
Male	Alive	Los Angeles	Medicaid	38
			Medicare	19
			Fee for Service	32
			PPO	11
			POS	57
			Managed Care	0
			HMO	0
			Private non-HMO	3
			Self Pay	6
			Local Charity	8
		Chicago	Medicaid	21
			Medicare	12
			Fee for Service	14
			PPO	22
			POS	47
			Managed Care	0
			HMO	0
			Private non-HMO	17
			Self Pay	11
			Local Charity	66
		New York	Medicaid	39
			Medicare	1
			Fee for Service	0
			PPO	24
			POS	144
			Managed Care	0
			HMO	0
			Private non-HMO	1
			Self Pay	1
			Local Charity	1
		Seattle	Medicaid	5
			Medicare	8
			Fee for Service	4
			PPO	5
			POS	33
			Managed Care	0
			HMO	0
			Private non-HMO	13
			Self Pay	3
			Local Charity	22
		Miami	Medicaid	6
			Medicare	8
			Fee for Service	12
			PPO	14
			POS	51
			Managed Care	0
			HMO	0
			Private non-HMO	0
			Self Pay	0

2155-0182/10/$3.00

		Local Charity	32
	Nashville	Medicaid	0
		Medicare	2
		Fee for Service	1
		PPO	0
		POS	0
		Managed Care	69
		HMO	0
		Private non-HMO	0
		Self Pay	0
		Local Charity	0
	Phoenix	Medicaid	2
		Medicare	4
		Fee for Service	3
		PPO	7
		POS	0
		Managed Care	24
		HMO	0
		Private non-HMO	8
		Self Pay	4
		Local Charity	2
Deceased	Los Angeles	Medicaid	5
		Medicare	2
		Fee for Service	1
		PPO	0
		POS	2
		Managed Care	0
		HMO	0
		Private non-HMO	0
		Self Pay	0
		Local Charity	0
	Chicago	Medicaid	0
		Medicare	0
		Fee for Service	0
		PPO	3
		POS	3
		Managed Care	0
		HMO	0
		Private non-HMO	1
		Self Pay	1
		Local Charity	4
	New York	Medicaid	6
		Medicare	1
		Fee for Service	0
		PPO	1
		POS	25
		Managed Care	0
		HMO	0
		Private non-HMO	0
		Self Pay	0
		Local Charity	0
	Seattle	Medicaid	2
		Medicare	0
		Fee for Service	0
		PPO	0
		POS	4
		Managed Care	0
		HMO	0
		Private non-HMO	0
		Self Pay	0
		Local Charity	5
	Miami	Medicaid	0
		Medicare	0
		Fee for Service	1
		PPO	4
		POS	2
		Managed Care	0
		HMO	0
		Private non-HMO	0
		Self Pay	3
		Local Charity	5
	Nashville	Medicaid	0
		Medicare	0
		Fee for Service	0
		PPO	0
		POS	0
		Managed Care	0
		HMO	8
		Private non-HMO	7
		Self Pay	0
		Local Charity	0
	Phoenix	Medicaid	1
		Medicare	0
		Fee for Service	1
		PPO	0
		POS	0
		Managed Care	5
		HMO	0
		Private non-HMO	2
		Self Pay	0
		Local Charity	0

Computing the total number of cells in a *cross-tabulation* of all categorical data as seen in Table 13 requires obtaining the product of the number of response categories for all variables. Status and gender both have 2 response categories, city has 7, and insurance has 10, so a total of 2x2x7x10=280 cells exist in Table 13.

Computing the total number of cells in a cross-tabulation of categorical *attributes* for a statistical analysis (the "design matrix") requires obtaining the product of the number of response categories for all attributes. For example, when predicting patient status using gender, city and insurance as attributes, the design matrix has a total of 2x7x10=140 cells. And, when predicting patient gender using status, city and insurance as attributes, the design matrix similarly has 140 cells. If observations were distributed uniformly in the cells (the opposite is in fact true), then on average 11.2 observations would exist in every cell of the design matrix.

When a linear analysis is conducted all categorical attributes having three of more response categories are usually reduced to a set of one-fewer binary dummy-coded indicator varia-

2155-0182/10/$3.00

bles than there are response options for the categorical scale.[5,6] Here, for example, city would be reduced to 6 binary indicators, and insurance to 9. To predict status or gender using the indicator variables instead of original city and insurance, implies a design matrix with 2 [gender or status] x (2x2x2x2x2x2) [city] x (2x2x2x2x 2x2x2x2x2) [insurance]=2x64x512, or a total of 65,536 cells (easily computed as $2^1x2^6x2^9=2^{16}$). Not only would the cross-classification table be *long* (each cell constitutes a row in the table), it would be *wide*. Displaying the cross-classification of these data would require 18 columns in Table 13, instead of the 5 columns used presently. If observations were distributed uniformly in the cells, then on average 0.024 observations would exist in every cell of the design matrix. Equivalently, there would be one observation for every 41.7 cells: a sparsely-populated table. This implies that most cells have n=0. Depending on the brand of statistical software used, substituting 0.5 for every empty cell would obviously add far more phantom subjects than in reality actually existed.

This analytic nightmare happens with only three categorical attributes included in the design. A rapid perusal of any academic journal reporting linear models for dichotomous class variables (dependent measures) will show that many studies employ numerous such attributes (independent variables) in their design.

An inherent, immitigable issue with of *all* so-called *suboptimal methods* is their failure to explicitly *maximize classification accuracy* (ESS) obtained by the model for a sample. Any model that explicitly returns maximum ESS for a sample is known as an *optimal* (or "maximum-accuracy") model, and any model unable to be proven to yield maximum ESS, but specifically engineered to seek maximum-accuracy solutions for a sample, is known as a *heuristic* maximum-accuracy model.[2] Inherent, immitigable issues for *all linear methods* are large size, small cell *n*, presence of numerous structural zeros (cells having n=0) in the design matrix, and the non-normality of the design matrix.

A thorough review of what is known as the *optimal data analysis* or ODA "maximum-accuracy" statistical analysis paradigm lies outside this article.[2,7] However, the issues presented by present data for suboptimal/linear methods vanish for ODA methods.[2,7] This is because, in contrast, *all* optimal methodologies—such as UniODA[2] used in univariate statistical analyses, and automated hierarchically optimal classification tree analysis[8,9] (CTA) methodology used in non-linear maximum-accuracy multivariate statistical analysis presented ahead—are specifically engineered to obviate these issues as well as a host of other issues that relate to use of the "conventional" statistical methodologies.[2,5-8]

Predicting patient status. Automated CTA was used to predict patient status (the class variable) using gender, city, and insurance as categorical attributes, with the following code:

```
VARS gender status city insure;
CLASS status;
ATTR gender city insure;
CATEGORICAL gender city insure;
MC ITER 5000 CUTOFF .05 STOP 99.9;
PRUNE .05;
ENUMERATE;
GO;
```

Results revealed no multiattribute CTA model was possible for these data, and the best solution identified was identical to the UniODA range-test solution for city, yielding the confusion table in Table 3. This finding is consistent with univariate results, which showed significant differences in status attributable only to city, and not to gender or type of insurance.

Predicting patient gender. Automated CTA was next used to predict patient gender (class variable) using status, city, and insurance as categorical attributes, by appending this code:

```
CLASS gender;
ATTR status city insure;
```

2155-0182/10/$3.00

```
CATEGORICAL status city insure;
GO;
```

A 3-attribute-based 4-segment partition of the sample was identified by CTA, yielding moderate accuracy (ESS=32.3) and weak predictive value (ESP=23.1). Figure 1 presents an illustration of the resulting CTA model. As is seen, CTA models initiate with a *root node*, from which two or more *branches* emanate and lead to other *nodes*: branches indicate pathways through the tree, and all branches terminate in model *endpoints*. The CTA algorithm chains together UniODA analyses in a procedure that explicitly identifies the combination of attribute subset and geometric structure that together predict the class variable with maximum possible accuracy (ESS) for the total sample.[8]

CTA models are highly intuitive: model "coefficients" are cutpoints or category descriptions expressed in their natural measurement units, and sample stratification unfolds in a flow process which is easily visualized across model attributes. Circles represent nodes in schematic illustrations of CTA models, arrows indicate branches, and rectangles represent model endpoints. Numbers (ordered attributes) or words (categorical attributes) adjacent to arrows give the value of the *cutpoint* (*category*) for the node. Numbers under nodes give the *experimentwise* Type I error rate for the node (in most research *estimated p* is reported). The number of observations classified into each endpoint is indicated under the endpoint and the percentage of targeted (here, female) observations is given inside the rectangle representing the endpoint.

Using CTA models to classify individual observations is straightforward. Imagine a hypothetical person on managed care living in LA. Starting at the root node, since the person lives in LA the left branch is appropriate. At the second node the right branch is appropriate because the person has managed care. Finally, at the third node the left branch is appropriate since the person is from LA. The person is thus classified into the corresponding model end-point: as seen, 11.7% of the observations classified into this model endpoint were females. Note that end-points represent sample strata identified by the CTA model. The probability of being female for this endpoint is $p_{female} \leq 0.117$: had the person instead lived in Chicago, then the right-hand endpoint would be appropriate, with $p_{female} \leq 0.240$.

Figure 1: CTA Model Predicting Gender

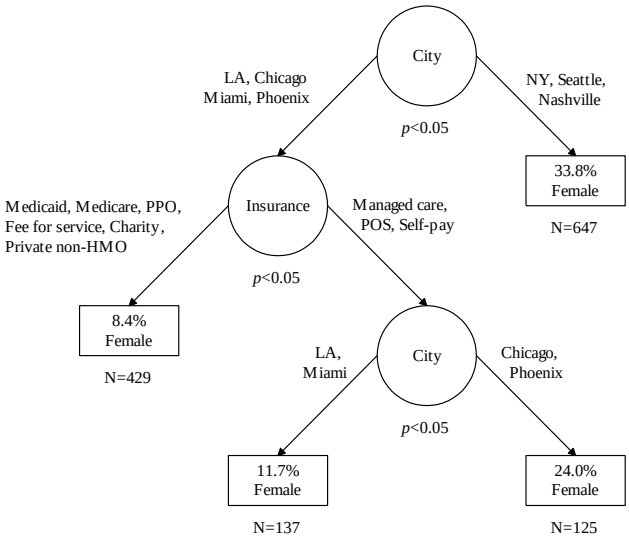

Table 14 presents the confusion table for the overall model.

Table 14: Confusion Table for CTA Model Predicting Gender Based on Patient Gender, Status and Insurance

		Predicted Gender		
		Male	Female	
Actual	Male	514	523	49.6%
Gender	Female	52	249	82.7%
		90.8%	32.2%	

The CTA model accurately classified the women in the sample, and it was accurate when it predicted a specific observation was a women. Therefore the model reflects the actual status of

2155-0182/10/$3.00

women well, but men presented a more complex profile—due in part to their larger numbers. The similarities and differences between univariate and multivariate findings are now considered.

For predicting patient *status* in UniODA analysis no statistically significant effects were obtained for gender or insurance, and neither of these attributes appeared in the CTA model (this pattern does not always occur). Presently CTA found the identical effect predicting status that UniODA identified when using city as attribute.

Predicting patient *gender* with UniODA there was no effect for status, and status did not appear in the CTA model. With UniODA there were significant effects found for city (based on final range-test: ESS=5.8; ESP=4.3) as well as for insurance (based on the final *experimentwise* range-test: ESS=17.9; ESP=12.6). Both of these attributes were included in the CTA model (this pattern does not always occur). City emerged as the most influential attribute in the CTA model, involved in the classification decisions for all of the observations in the sample, while insurance was only involved in classifications of *n*=1,568-647 observations (see Figure 1)—corresponding to 58.7% of the total sample.

Note that the CTA-based order of cities with respect to percent of females in the model endpoints is: (LA, Miami)<(Chicago, Phoenix)<(NY, Seattle, Nashville). This is identical to the UniODA model in the second step of the range test (Table 7), but it is not the final model obtained by UniODA for city (Table 8): additional reduction as occurred in UniODA would reduce overall ESS of the CTA model (this same argument may be used to select the higher-ESS UniODA model identified earlier). Considering insurance groupings parameterizing CTA model branches (Figure 1), UniODA and CTA model left-hand branches shared Medicare, and right-hand branches shared managed care: the other insurance types were all on the *opposite* branch, and the CTA model did not include HMO in the roster of insurance categories (HMO was not an insurance category for the cities in the left-hand branch of the CTA model emanating from the root node; Figure 1). This illustrates very well the difference between UniODA and CTA: the former finds the optimal (maximum ESS) solution for the sample considering one attribute at a time in isolation of all other attributes; the latter finds the optimal (maximum ESS) solution for the sample considering all attributes included in analysis in conjunction with one another. Table 15 is the CTA staging table[8].

Table 15: Staging Table for CTA Model Results

Stage	City	Insure	City	*n*	p_{female}	Odds
1	LA, Chicago, Miami, Phoenix	Medicaid, Medicare, Fee for Service, PPO, Private non-HMO, Charitable group	---	429	0.084	1:11
2	LA, Chicago, Miami, Phoenix	POS, Managed care, Self-pay	LA, Miami	137	0.117	1:8
3	LA, Chicago, Miami, Phoenix	POS, Managed care, Self-pay	Chicago, Phoenix	125	0.240	1:3
4	NY, Seattle, Nashville	---	---	647	0.338	1:2

2155-0182/10/$3.00

Staging tables are an intuitive alternative representation of CTA findings, useful for defining "propensity" scores (weights) to assign to all observations based on the findings of the CTA model.[1] The rows of the staging table are the model end-points reorganized in increasing order of percent of class 1 (female) membership. Stage is thus an *ordinal index* of propensity, and p_{female} is a *continuous index*: increasing values on either index indicates increasing propensity. Compared to Stage 1, p_{female} is 1.4-times greater in Stage 2; 2.9-times greater in Stage 3; and 4.0-times greater in Stage 4.

To use the table to stage a given observation, simply evaluate the fit between the observation's data and each stage descriptor. Begin at Stage 1, and work sequentially through stages until identifying the descriptor which is *exactly true* for the data of the observation undergoing staging. Consider the hypothetical person discussed earlier living in LA with managed care. Starting with Stage 1, city is appropriate, but insurance does not include managed care. Moving to Stage 2, city is appropriate (LA), insurance is appropriate (managed care), and the second city column is appropriate (LA): the person is thus classified as Stage 2 along with 136 other people in the sample. The Stage 2 patient strata is 11.7% female: odds of being female in Stage 2 are thus 1:8.

If the numerator of the presented odds is one, then the denominator of the presented odds is $(1/\underline{p_{female}})$-1. For example, for Stage 1 $\underline{p_{female}}$ is 0.0884, so denominator=(1/0.084)-1=11.905-1, or 10.91. In Table 15 the odds for Stage 1 are given as 1:11.

The CTA model achieved greater overall ESS and ESP than any of the UniODA models; greater sensitivity in accurately classifying the actual women in the sample than any of the UniODA models; and was surpassed in ability to make accurate classifications of observations as being women by one UniODA model (Table 7). The CTA model segmented the sample into four partitions: this level of discrimination gradation was only achieved by the UniODA model for predicting gender based on city.

ESS and ESP index the overall strength of the model. Model efficiency, computed as the mean strength index divided by the number of sample partitions (segments) that are identified by the model, adjusts classification performance to reflect relative complexity (complexity is the opposite of parsimony).[2] For ESS the efficiency for the final UniODA and CTA models are 1.4 and 8.1, and for ESP are 1.1 and 5.8 respectively, so the CTA indices are 479% and 427% higher than corresponding UniODA indices, respectively.

However, as mentioned earlier, it may be argued that the optimal model for discriminating gender based on city via UniODA was the initial model with two endpoints, for which ESS=31.5 and ESP=22.0 (Table 6). This is the strongest of the UniODA models in this analysis and also the most parsimonious: with two endpoints mean ESS and ESP for this UniODA model are thus 15.8 and 11.0. These latter mean values are 95% and 90% *greater* than were achieved using the CTA model. Thus, when considering the crucial role of parsimony in theory development (which is the primary function of the ESS statistic[3]), the insurance attribute halves the models efficiency (two versus four sample partitions), in exchange for a modest gain in ESS (32.3 for CTA versus 31.5 for UniODA) and ESP (23.1 versus 22.0). Seen in this light, city facilitates a moderate level of accuracy and weak predictive value for discriminating gender, and the type of insurance does *not* increase discrimination to a practically significant degree.

References

[1]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmiel J, Sipler A, Chan C, Goetz MB, Schwartz D, Bennett CL (2000). A new preadmission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respir-*

2155-0182/10/$3.00

atory and Critical Care Medicine, *161*, 1081-1086.

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

[3]Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119.

[4]Yarnold JK (1970). The minimum expectation of χ^2 goodness-of-fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Society*, *65*, 864-886.

[5]Grimm LG, Yarnold PR (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, D.C.: APA Books, 1995.

[6]Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: APA Books, 2000.

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

[8]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160.

[9]Yarnold PR (2013). Initial use of hierarchically optimal classification tree analysis in medical research. *Optimal Data Analysis*, 2, 7-18.

Author Notes

E-mail: Journal@OptimalDataAnalysis.com

ODA Blog: http://odajournal.com