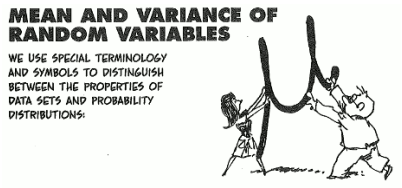
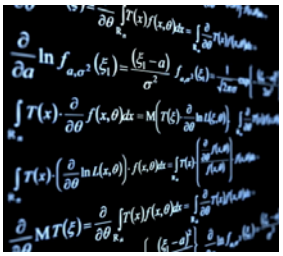


Information Theory and Model Selection



Some notes

http://video.sdsu.edu/nas/capture/2015/ddeutschman/Information_Theory/Information_Theory_-_20150216_132003_9.mp4

Some references

- General: Discussing IT approaches relative to classical hypothesis testing (and Bayesian)
  - Beninger et al. 2012. Strengthening statistical usage in marine ecology.
  - Garamszegi et al. 2009. Changing philosophies and tools for statistical inferences in behavioral ecology.
  - Hobbs and Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: A guide to self teaching
- IT approaches including model selection and multimodel approaches
  - Anderson and Burnham. 2014. Multimodel Inference Understanding AIC and BIC in Model Selection
  - Anderson et al. 2012. AIC Model selection and multimodel inference: some background, observations, and comparisons.
  - Symonds and Moussalli. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion

Some more references

- AIC, AICc and BIC
  - Aho et al. 2014. Model selection for ecologists: the worldviews of AIC and BIC.
  - Ward. 2008. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools.
- Pros, Cons, suggestions, and cautions
  - Murtaugh. 2009. Performance of several variable-selection methods applied to real ecological data.
  - Mundry. 2011. Issues in information theory-based statistical inference - a commentary from a frequentist's perspective.
  - Galipaud et al. 2014. AIC Model selection and multimodel inference: some background, observations, and comparisons.
  - Grueber et al. 2011. Multimodel inference in ecology and evolution: challenges and solutions.

Opinion TRENDS in Ecology and Evolution Vol.22 No.4 ScienceDirect

Inference in ecology and evolution

Philip A. Stephens[1], Steven W. Buskirk[2] and Carlos Martínez del Río[2]
[1]Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK
[2]Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071-3166, USA

Introduction:
biologists study complex systems that are characterized by high natural variability and, not surprisingly, rely heavily on statistics to infer pattern and causation from their data. In that context, the use of null hypothesis significance tests (NHST) predominates. . . . Over the past decade, several biologists have emphasized the limitations and problems of NHST, promoting a variety of alternatives In spite of this, NHST remains the main approach to inference in ecology and evolution.

Abstract:
Most ecologists and evolutionary biologists continue to rely heavily on null hypothesis significance testing, rather than on recently advocated alternatives, for inference. Here, we briefly review null hypothesis significance testing and its major alternatives. We identify major objectives of statistical analysis and suggest which analytical approaches are appropriate for each. Any well designed study can improve our understanding of biological systems, regardless of the inferential approach used. Nevertheless, an awareness of available techniques and their pitfalls could guide better approaches to data collection and broaden the range of questions that can be addressed . . .

Box 1. Current use of statistical approaches in ecology and evolution

We reviewed the last 50 empirical papers published in 2005 in each of four journals in ecology and evolutionary biology (*Behavioral Ecology*, *Ecology Letters*, *Evolution* and the *Journal of Applied Ecology*). Reviews and purely modelling-based papers were excluded, as we were principally interested in how inferences were drawn from data.

Papers were scored according to whether they used NHST (black bars), information theoretic approaches (IT; red bars), or other approaches (principally Bayesian analyses or likelihood-based phylogenetic tree constructions; green bars). Some papers used more than one type of approach and, thus, totals sum to more than 100%. In all four journals, most papers (≥90% in each case) used NHST, whereas ≤10% used IT (Figure I). Other techniques were commonly used only in *Evolution*, which is perhaps unsurprising, given the relative frequency with which authors in that journal deal with phylogenetic inferences.

It has been suggested that NHST approaches can be appropriate in experimental studies but should not be used in observational studies (because variance in the data set has not been generated by experimental manipulation, leaving inference vulnerable to unconsidered confounding factors) [6]. Consequently, we also scored papers according to whether they used observational or experimental data (*Behavioral Ecology*, $n = 21$; *Ecology Letters*, $n = 26$; Evolution, $n = 31$; *Journal of Applied Ecology*, $n = 38$), and assessed the frequency with which NHST approaches were used (blue bars). These frequencies tended to be only marginally lower than the frequencies with which NHST was used overall (Figure I).

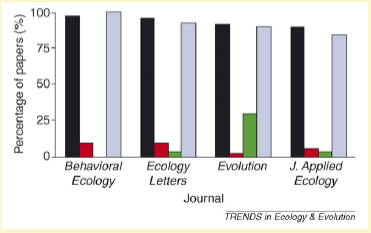


Figure I.

Table 1. An overview of inferential approaches

Approach	Requirements	Outcomes	Advantages	Disadvantages
Null hypothesis significance testing	Data, **Y**, and a statistical null hypothesis, H_0, which designates the test statistic of interest, t	Provides $P(t \geq t_{obs}\|H_0)$, the probability of observing the test statistic (or one more extreme), if the null hypothesis is true. In carefully designed and well replicated experiments, NHST enables H_A, the converse of the null, to be falsified (if experiments repeatedly fail to reject the null with a suitably low P value)	Computational simplicity (with ready availability of user interfaces)	A variety of inherent difficulties with interpretation, as well as deeper philosophical problems that can limit scientific advances
Information theoretic model comparison	Data, **Y**, and a set of two or more competing models, $H_1 \ldots H_n$, which might include the null and its converse, a nested set of arrangements of potential predictor variables, or several disparate, mechanistic descriptions of a system	Provides an information criterion value for each competing model, usually of the form $C = -2\ln[L(H_i\|\mathbf{Y})] + B$, where C is the criterion estimate, $L(H_i\|\mathbf{Y})$ is the model likelihood and B is a penalty imposed by some aspect of the model or data (e.g. Ref. [5])	Enables models to be ranked in order of relative support; evidence ratios to be calculated for any pair of competing models; and model averaging to account for uncertainty in model selection ([6,17,27] but see Ref. [55]). Discourages binary approaches to inference	Unclear which information criterion is most appropriate [26,27] or how well some criteria perform under different conditions [55]
Bayesian statistics	Data, **Y**, a set of competing models (as above) and prior information, which might include previous estimates of the plausibility of each model, as well as their parameter values	Unique in providing $P(H_i\|\mathbf{Y})$ [4]	Including existing knowledge means that knowledge accumulates; sensitivity analyses are intrinsically accommodated (through presenting posteriors for a range of prior assumptions); all uncertainties are integrated out; and Bayesian approaches can deal with complex problems, such as those with both process and observation errors	Computational complexity, as Bayesian approaches require integrating under likelihood functions; this can render even a simple ANOVA complex in a Bayesian framework [56]
Effect size statistics	Data, **Y**, from two or more treatment groups	Measures of the practical significance of an observed effect (e.g. the difference between two means), e.g. the counternull [21], Cohen's d[a] [57], the CL statistic of McGraw and Wong[b] [58,59], and several other measures [58,60]	Focuses attention away from statistical significance and resultant biases; improves the potential for meta-analyses of experimental data (but see Ref. [61])	Effect size statistics are largely descriptive and, as such, are often unsatisfactory as sole measures of the outcome of an experiment

[a] A standardized descriptor quantifying, independently of sample size, the degree to which the means of two treatment groups are separated.
[b] An estimate of the probability that a randomly chosen subject in one treatment class will have a higher value of the test statistic than a random subject from another treatment

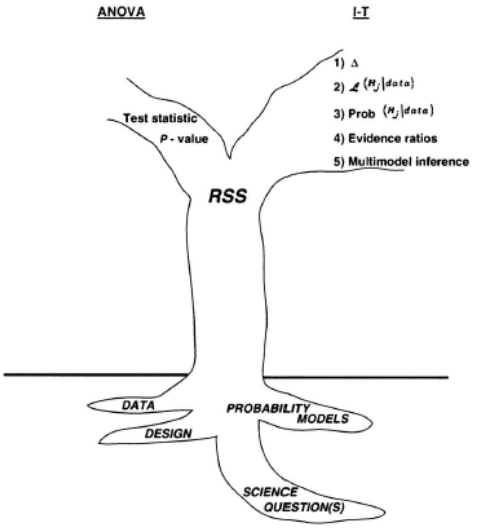


Fig. 3 I-T approaches provide a superior alternative to the traditional test statistic and P value paradigm (e.g., t tests and ANOVA tables). A conceptual diagram of the pivotal branching point in the tree is the RSS. All the important issues that precede data analysis (the roots) are the same under either analysis approach, including the estimated effect size and its precision for a given model

Behav Ecol Sociobiol (2011) 65:13–21
DOI 10.1007/s00265-010-1037-6

REVIEW

A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion

Matthew R. E. Symonds · Adnan Moussalli

AIC is calculated as

$$\mathrm{AIC} = -2\ln(L) + 2k$$

if using likelihood or

$$\mathrm{AIC} = n\left[\ln\left(\frac{\mathrm{RSS}}{n}\right)\right] + 2k$$

if using residual sum of squares, where n is the sample size.

For small sample sizes (roughly approximated as being when n/k is less than 40 and k is the number of fitted parameters in the most complex model), a modified version of AIC ($\mathrm{AIC_c}$) is recommended:

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

In practice, because $\mathrm{AIC_c}$ approximates AIC at large sample sizes, it is often advised that $\mathrm{AIC_c}$ is used as default

Table 1 Major statistical packages and how they implement Akaike's information criterion

Package	Website	Versions of AIC calculated	Additional Notes
Genstat 12	www.vsni.co.uk/software/genstat/	AIC only—automatically calculated for generalised linear models and for restricted maximum likelihood (REML) for linear mixed models	For generalised linear models, the model variance is not taken into account in the count of fitted parameters; for REML, the variance parameters in the random model are included in the parameter count
JMP 8	www.jmp.com	$\mathrm{AIC_c}$ only—automatically calculated when analysing in the stepwise regression menu	Stepwise regression menu also allows one to compare $\mathrm{AIC_c}$ across all possible models, calculate Akaike weights and perform model averaging
Minitab 15	www.minitab.com	None	
SPSS 18	www.spss.com/statistics/18/	AIC and $\mathrm{AIC_c}$—calculated in several procedures (e.g. generalised linear models, mixed models, time series analysis)	For generalised linear models, the model variance is included in the parameter count
R	http://www.r-project.org/	AIC only, in two commands: extractAIC {stats} &AIC {stats}	For full details, see http://stat.ethz.ch/R-manual/R-patched/library/stats/html/AIC.html
SAS 9.2	www.sas.com	AIC and $\mathrm{AIC_c}$ supplied as part of the 'Fit Statistics' table in numerous procedures	For simple and generalised linear models, the model variance is not taken into account in the count of fitted parameters, but it is included for the generalised linear mixed model procedure. AICc is only provided under the generalised linear and mixed model applications. Number of parameters is explicitly stated in the output
Statistica 9	www.statsoft.com	AIC only—calculated as part of regression output	Provides an option to report AIC for all possible models. Output shows how many parameters are used
Systat 12	www.systat.com	AIC and $\mathrm{AIC_c}$ supplied as part of output of numerous procedures	

3. UNDERSTANDING BIC

Schwarz (1978) derived the Bayesian information criterion as

$$\mathrm{BIC} = -2\ln(\mathcal{L}) + K\log(n).$$

As usually used, one computes the BIC for each model and selects the model with the smallest criterion value. BIC is a misnomer as it is not related to information theory. As with $\Delta\mathrm{AIC}_i$, we define $\Delta\mathrm{BIC}_i$ as the difference of BIC for model g_i and the minimum BIC value. More complete usage entails computing posterior model probabilities, p_i, as

$$p_i = \Pr\{g_i|\mathrm{data}\} = \frac{\exp(-\frac{1}{2}\Delta\mathrm{BIC}_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta\mathrm{BIC}_r)}$$

Model selection procedures in social research: Monte-Carlo simulation results

Lawrence E. Raffalovich*, Glenn D. Deane, David Armstrong and Hui-shien Tsao

Journal of Applied Statistics
Vol. 35, No. 10, October 2008, 1093–1114

Abstract:
Model selection strategies play an important, if not explicit, role in quantitative research. The inferential properties of these strategies are largely unknown, therefore, there is little basis for recommending (or avoiding) any particular set of strategies. In this paper, we evaluate several commonly used model selection procedures [Bayesian information criterion (BIC), adjusted *R*2, Mallows'*C*p, Akaike information criteria (AIC), AICc, and stepwise regression] using Monte-Carlo simulation of model selection when the true data generating processes (DGP) are known.

We find that the ability of these selection procedures to include important variables and exclude irrelevant variables increases with the size of the sample and decreases with the amount of noise in the model. None of the model selection procedures do well in small samples, even when the true DGP is largely deterministic; thus, data mining in small samples should be avoided entirely. Instead, the implicit uncertainty in model specification should be explicitly discussed. In large samples, BIC is better than the other procedures at correctly identifying most of the generating processes we simulated, and stepwise does almost as well. In the absence of strong theory, both BIC and stepwise appear to be reasonable model selection strategies in large samples. Under the conditions simulated, adjusted *R*2, Mallows' *C*p AIC, and AICc are clearly inferior and should be avoided.

Y: state crime rate in 1960
X_1: percent of males, 14–24
X_2: southern state
X_3: mean years of schooling
X_4: police expenditure in 1960
X_5: police expenditure in 1959
X_6: labor force participation rate
X_7: males per 100 females
X_8: state population
X_9: non-whites per 1000 population
X_{10}: unemployment rate of urban males, 14–24
X_{11}: unemployment rate of urban males, 35–39
X_{12}: gross domestic product (GDP)
X_{13}: income inequality
X_{14}: probability of imprisonment
X_{15}: average time served in state prison

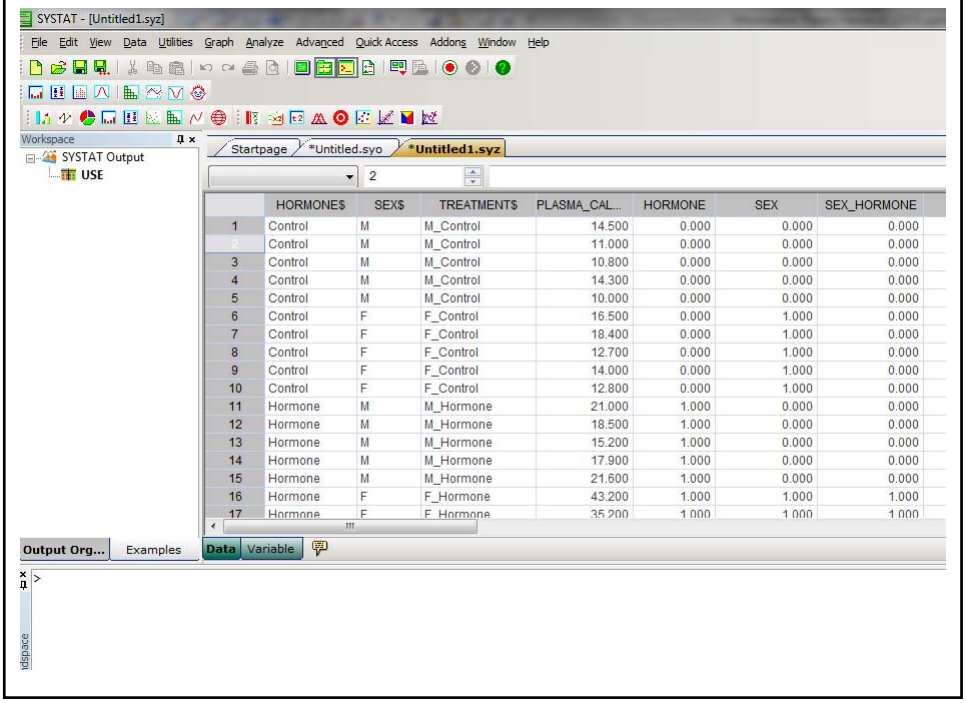
We use these data to specify the six linear models presented and discussed by Raftery [15, pp. 120–124]:

Model 1: $Y_1 = f(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, U_1)$
Model 2: $Y_2 = f(X_1, X_3, X_4, X_9, X_{11}, X_{13}, X_{14}, U_2)$
Model 3: $Y_3 = f(X_1, X_2, X_4, X_9, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, U_3)$
Model 4: $Y_4 = f(X_1, X_3, X_4, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, U_4)$
Model 5: $Y_5 = f(X_3, X_{12}, X_{13}, X_{14}, X_{15}, U_5)$
Model 6: $Y_6 = f(X_1, X_4, X_9, X_{10}, X_{12}, X_{13}, X_{14}, X_{15}, U_6)$

Table 4. Summary of most successful selection procedures.

	Correct model $\sigma = 0.74$	Correct model $\sigma = 0.41$	Correct model $\sigma = 0.18$
$N = 47$			
Model 1	None	None	Adj. R^{2*}
Model 2	None	AIC*	Stepwise*
Model 3	None	Adj. R^{2*}	C_p^*
Model 4	None	None	Adj. R^{2*}
Model 5	C_p, AIC*	AIC_c^*	Stepwise
Model 6	None	None	C_p^*
$N = 100$			
Model 1	None	None	Adj. R^{2*}
Model 2	Adj. R^{2*}	AIC*	Stepwise*
Model 3	None	Adj. R^{2*}	Stepwise**
Model 4	None	Adj. R^{2*}	AIC*
Model 5	AIC_c^*	BIC*	BIC
Model 6	None	Adj. R^{2*}	AIC_c^*
$N = 500$			
Model 1	None	Adj. R^{2*}	Adj. R^{2*}
Model 2	C_p^*	Stepwise**	BIC
Model 3	Adj. R^{2*}	Stepwise*	BIC
Model 4	Adj. R^{2*}	C_p, AIC_c^*	BIC
Model 5	BIC**	BIC	BIC
Model 6	C_p, AIC, AIC_c^*	Adj. R^{2*}	BIC*
$N = 1000$			
Model 1	Adj. R^{2*}	Adj. R^{2**}	Adj. R^{2**}
Model 2	BIC	BIC	BIC
Model 3	BIC	BIC	BIC
Model 4	Stepwise	BIC	BIC
Model 5	BIC	BIC	BIC
Model 6	Stepwise*	Stepwise**	BIC

Note. *Success rate <25%, **Success rate <50%.



Model Weights and Model Averaging

$$\mathrm{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+1)}{n-K-1}$$

(see Sugiura 1978; Hurvich and Tsai 1989, 1995), and this should be used unless $n/K >$ about 40 for the model with the largest value of K. A pervasive mistake in the model selection literature is the use of AIC when AIC_c really should be used. Because AIC_c converges to AIC, as n gets large, in practice, AIC_c should be used. People often conclude that AIC overfits because they failed to use the second-order criterion, AIC_c.

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min},$$

where $\mathrm{AIC}_{\min}$ is the minimum of the R different AIC_i values (i.e., the minimum is at $i = \min$). This transformation forces the best model

AKAIKE WEIGHTS, w_i

It is convenient to normalize the model likelihoods such that they sum to 1 and treat them as probabilities; hence, we use

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R} \exp(-\Delta_r/2)}.$$

The w_i, called Akaike weights, are useful as the "weight of evidence" in favor of model $g_i(\cdot|\theta)$ as being the actual K-L best model in the set (in this context, a model, g, is considered a "parameter").

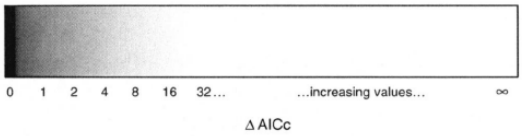


Fig. 2 Plausible hypotheses are identified by a narrow region in the continuum where Δ< perhaps four to seven (*black and dark grey*). The evidence in the light grey area is inconclusive and value judgments for hypotheses in this region are equivocal. Implausible models are shown in *white*, Δ> about 14

The averaged parameter estimate is calculated as follows:

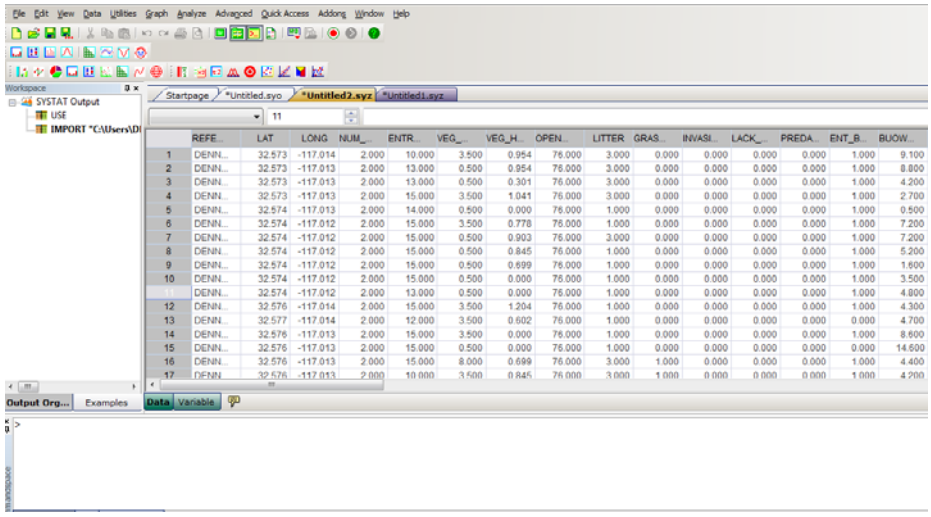
$$\bar{\hat{\beta}} = \frac{\sum_{i=1}^{R} w_i \hat{\beta}_i}{\sum_{i=1}^{R} w_i}$$

where $\hat{\beta}_i$ is the estimate for the predictor in a given model i, and w_i is the Akaike weight of that model. In this

variance (an estimate of variance not conditional on a single model) can then be calculated using the following equation:

$$\widehat{var}\left(\bar{\hat{\beta}}\right) = \left[\sum w_i \sqrt{\widehat{var}\left(\hat{\beta}_i\right) + \left(\hat{\beta}_i - \bar{\hat{\beta}}\right)^2}\right]^2$$

where $\widehat{var}\left(\hat{\beta}_i\right)$ is the variance of the parameter estimate in model i, and $\hat{\beta}_i$ and $\bar{\hat{\beta}}$ are as defined above (Buckland et al. 1997). Note that AIC weights for the subset of models for which the variate of interest appears need to be renormalized (i.e. summed up to 1) before calculating the


File Edit View Data Utilities Graph Analyze Advanced Quick Access Addons Window Help
Workspace
SYSTAT Output
USE
IMPORT "C:\Users\D
Startpage
*Untitled.syo
*Untitled2.syz
*Untitled1.syz
11
REFE... LAT LONG NUM_... ENTR... VEG_... VEG_H... OPEN... LITTER GRAS... INVASI... LACK_... PREDA... ENT_B... BUOW...
1 DENN... 32.573 -117.014 2.000 10.000 3.500 0.954 76.000 3.000 0.000 0.000 0.000 0.000 1.000 9.100
2 DENN... 32.573 -117.013 2.000 13.000 0.500 0.954 76.000 3.000 0.000 0.000 0.000 0.000 1.000 8.800
3 DENN... 32.573 -117.013 2.000 13.000 0.500 0.301 76.000 3.000 0.000 0.000 0.000 0.000 1.000 4.200
4 DENN... 32.573 -117.013 2.000 15.000 3.500 1.041 76.000 3.000 0.000 0.000 0.000 0.000 1.000 2.700
5 DENN... 32.574 -117.013 2.000 14.000 0.500 0.000 76.000 1.000 0.000 0.000 0.000 0.000 1.000 0.500
6 DENN... 32.574 -117.012 2.000 15.000 3.500 0.778 76.000 1.000 0.000 0.000 0.000 0.000 1.000 7.200
7 DENN... 32.574 -117.012 2.000 15.000 0.500 0.903 76.000 3.000 0.000 0.000 0.000 0.000 1.000 7.200
8 DENN... 32.574 -117.012 2.000 15.000 0.500 0.845 76.000 1.000 0.000 0.000 0.000 0.000 1.000 5.200
9 DENN... 32.574 -117.012 2.000 15.000 0.500 0.699 76.000 1.000 0.000 0.000 0.000 0.000 1.000 1.600
10 DENN... 32.574 -117.012 2.000 15.000 0.500 0.000 76.000 1.000 0.000 0.000 0.000 0.000 1.000 3.500
11 DENN... 32.574 -117.012 2.000 13.000 0.500 0.000 76.000 1.000 0.000 0.000 0.000 0.000 1.000 4.800
12 DENN... 32.576 -117.014 2.000 15.000 3.500 1.204 76.000 1.000 0.000 0.000 0.000 0.000 1.000 4.300
13 DENN... 32.577 -117.014 2.000 12.000 3.500 0.602 76.000 1.000 0.000 0.000 0.000 0.000 0.000 4.700
14 DENN... 32.576 -117.013 2.000 15.000 3.500 0.000 76.000 1.000 0.000 0.000 0.000 0.000 1.000 8.600
15 DENN... 32.576 -117.013 2.000 15.000 0.500 0.000 76.000 1.000 0.000 0.000 0.000 0.000 0.000 14.600
16 DENN... 32.576 -117.013 2.000 15.000 8.000 0.699 76.000 3.000 1.000 0.000 0.000 0.000 1.000 4.400
17 DENN... 32.576 -117.013 2.000 10.000 3.500 0.845 76.000 3.000 1.000 0.000 0.000 0.000 1.000 4.200
Output Org...
Examples
Data
Variable

All Done