Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 2

11-1-2003

Model Comparisons Using Information Measures

C. Mitchell Dayton
University of Maryland, cdayton@umd.edu

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 281-292

INVITED ARTICLES

Model Comparisons Using Information Measures

C. Mitchell Dayton
University of Maryland

Methodologists have criticized the use of significance tests in the behavioral sciences but have failed to provide alternative data analysis strategies that appeal to applied researchers. For purposes of comparing alternate models for data, information-theoretic measures such as Akaike AIC have advantages in comparison with significance tests. Model-selection procedures based on a min(AIC) strategy, for example, are holistic rather than dependent upon a series of sometimes contradictory binary (accept/reject) decisions.

Key words: Akaike AIC, significance tests, information measures

Introduction

Quantitative researchers have been trained to evaluate effects of interest utilizing the methods of statistical inference. In a single research study it is not unusual to see several dozen, or even several hundred, significance tests applied to assess, for example, multiple correlations, differences among multiple correlations and regression coefficients. However, the appropriateness of the use of significance tests in social and behavioral research settings has been debated for more than 40 years. In particular, Rozeboom (1960) summarized criticisms of significance testing that have resurfaced in various guises from time to time. Generally, these criticisms have focused on the issue of binary decision-making (e.g., accept/reject null hypotheses) as opposed to considerations related to weight of evidence (e.g., measures of strength of effect or effect sizes).

The fundamental error, as seen by Rozeboom, "…lies in mistaking the aim of a scientific investigation to be a decision, rather than a cognitive evaluation of propositions (op. cit., page 212)." Although distinctions can be drawn between significance testing in the Fisherian (1959) sense and hypothesis testing in the Neyman-Pearson (1933) sense, current teaching and practice in the behavioral sciences blur these distinctions and the terms can be considered as essentially interchangeable in practice. However, it is likely that Fisher himself would concur with many of the criticisms as suggested by the following quotes (Fisher, 1959):

> …the calculation {of significance levels} is absurdly academic, for in fact no scientific worker has a fixed

C. Mitchell Dayton is a Professor of Measurement & Statistics at the University of Maryland. His major research interests deal with the topics of latent class analysis and simultaneous inference. He recently published a Sage book dealing with latent class scaling models, a topic on which he has published widely. His long standing interest in simultaneous inference has led to a focus on model-comparison approaches utilizing information theory and posterior Bayes factors.

> level of significance at which from year to year, and in all circumstance, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (page 42)
>
> On the whole the ideas (a) that a test of significance must be regarded as one of a series of similar tests applied to a succession of similar bodies of data, and (b) that the purpose of the test is to discriminate or 'decide' between two or more hypotheses, have greatly obscured their understanding, when taken not as contingency possibilities but as elements essential in their logic. (page 42)

Advocates for change have urged minimizing (or, even eliminating) the role of significance tests in behavioral research and elevating the roles of procedures such as confidence intervals, measures of effect size (e.g., Carver, 1993) or replicability measures (e.g., Thompson, 1994). Although these advocacy positions have been well articulated and widely disseminated among applied statisticians, there is scant evidence for change in practice by applied researchers in the behavioral sciences.

For example, the Fall 1995, Winter 1995 and Spring 1996 issues of the American Educational Research Journal contained 11 data-based articles in the Teaching, Learning and Human Development section of the journal. The number of significance tests per article (with some allowances for counting errors) are, in rank order: 3, 29, 33, 35, 40, 48, 94, 212, 290, 335 and 448 for a total of 1567 tests or an average of 522 significance tests per issue of the journal.

Although the lowest number, 3, might lead to useful interpretations within a single research study, it is highly doubtful that 29, much less 448, such tests in a single study can be interpreted in manner that provides much scientific value. Indeed, the lack of popularity for alternative procedures to significance testing has, itself, a long history as evidenced by Heermann and Braskamp (1970) who wrote in the Introduction to Part 4, Testing Statistical Hypotheses, of their book of readings:

> there is considerable agreement among statisticians and behavioral scientists that there has been an unfortunate emphasis on the part of the latter on hypothesis testing to the exclusion of other inferential techniques….In spite of this widely known fact, behavioral scientists continue to employ significance tests to the exclusion of other more informative techniques. (page 154)

It can be argued that a major reason for the apparent resistance to change from significance tests to other techniques is that the alternatives that have been proposed are unattractive to applied researchers. Consider the relatively simple example of multiple comparisons among a set of, say, five sample means. A typical traditional approach would be the use of Tukey tests (or one of the plethora of variations such as Games-Howell tests). In effect, 10 significance tests would be conducted and referred to the appropriate theoretical distribution (e.g., studentized range).

If a researcher were to follow Carver's (1993) advice, the Tukey tests would be replaced by "…estimates of effect size and of sampling error such as standard errors and confidence intervals 89)." However, the q statistic per se can be viewed as an effect size (i.e., difference between two means divided by the estimated standard error of a mean) and how does the researcher arrive at a unified interpretation of the 10 confidence intervals? But Carver (1993) has additional advice: "Better yet, by conducting multiple studies, replication of results can replace statistical significance testing." This is not a particularly attractive option given the obstacles that may exist to replication and the fact that the researcher really needs to interpret the present study in order to decide whether or not replication is a worth while expenditure of time and resources.

A premise of this paper is that significance tests are appropriate for only certain, highly constrained purposes but have enjoyed much wider use because of the failure of methodologists to popularize other, more appropriate statistical methods. In particular, significance tests are useful for interpreting data that arise from controlled experimental or quasiexperimental designs in which the role of specific hypotheses is well-defined. For non-experimental settings, researchers typically utilize significance tests for purposes of comparing alternate models for data or for interpreting effects within specific models. It is this application that is better served by procedures specifically designed for comparisons among models and is ill-served by significance tests.

An increasingly popular technique for comparing models involves information-theoretic measures such as Akaike (1973, 1978) AIC or measures based on posterior Bayes factors such as Schwarz (1978) BIC. In either case, these measures may be viewed as penalized log-likelihoods and are computed separately for each model under consideration. Then, a preferred model, among those being compared, can be selected.

This permits a very wide range of applications and even avoids some technical issues in applying statistical tests for model comparisons (e.g., for comparing number of components for discrete mixture models such as latent class models). Model comparison procedures are holistic in the sense that a variety of competing models can be assessed simultaneously and a best model selected by applying a single rule. Attempting to compare models using significance tests is, by contrast, piece-meal with the final selection of a model based on results from sometimes conflicting outcomes.

Consider, for example, the procedure that is often used when fitting polynomial regression models to bivariate data. Assume there are five distinct levels of a quantitative independent variable, X, so that models corresponding to linear, quadratic, cubic and quartic regression can meaningfully be fit to the data. Typically, the differences in fit of increasingly more complex models are evaluated by significance tests based on differences in multiple correlations (of, equivalently, differences in explained variability).

Thus, four distinct hypotheses are tested with, say, four hierarchical F statistics each at some specified level of significance. Since four independent tests are being conducted, an initial decision is whether or not to control the Type I error rate for the set of tests or, simply, to use a conventional .05 level for each test. This decision, it should be noted, can dramatically affect the interpretation of results. On the other hand, a holistic, model-comparison approach entails computing, say, an Akaike AIC statistic for each regression model and then selecting a "best" model corresponding to the minimum value of AIC.

Another consideration in selecting an approach to comparing models is the logic of the decision-making strategy itself. In applying significance tests, the null hypothesis corresponds to some restricted form of a model (e.g., a test for quadratic regression involves a null hypothesis stating that the regression coefficient for the quadratic term is zero and this corresponds to a simpler, linear regression model). The validity of the test depends upon assuming that the simpler model is true and that deviations from the model are due to chance. But this is a gross over-simplification of the scientific process. In a holistic, model-comparison approach the underlying goal is to select the best approximating model from among the models under consideration. It is not necessary to assume that any given model is "true" and there is no need to posit that a true model exists among those being compared.

In this article, the rationale for information-theoretic model comparison procedures is presented and two specific areas of application are discussed – pairwise comparisons and analysis of finite mixtures.

Information Criteria

Akaike (1973) suggested that the Kullback-Leibler (1951) information measure provides a natural criterion for ordering alternate models for data. He developed a sample-based estimate, AIC, for this information measure that he incorporated into a decision-making strategy. For any specific model, the form of AIC is $-2LL+2p$ where LL is the log-likelihood for the model and p is the number of independent parameters that are estimated in fitting the model to data.

For example, assuming normally distributed residuals for a homogeneous linear regression model for three independent variables, p would equal five and comprise three partial regression coefficients, the mean of the dependent variable (or the Y-intercept) and the variance of the residuals.

A summary of the technical development for the AIC measure can be found in Dayton (2003a) whereas a detailed analysis of the measure is presented by de Leeuw (1992). In general terms, Kullback-Leibler information is a measure of the discrepancy between the true distribution for a random variable (possibly vector-valued) and the distribution specified by some particular model. Although the true model is never known, Akaike managed to derive an estimate of this discrepancy by considering the distribution of a future sample conditional on knowing the maximum-likelihood estimator for parameters in the model.

Fundamentally, AIC involves the notion of cross-validation, but only in a theoretical sense. Given AIC values for two or more alternate models, the model satisfying min(AIC) is, in this information-theoretic sense, most representative of the true model and, on this basis, may be interpreted as the best approximating model among those being considered. A useful interpretation of AIC is that it estimates the loss of precision (or, increase in information) that results from substituting maximum likelihood estimates for the true parametric values in the likelihood function. Thus, among the models under consideration, it can be argued that the preferred model (i.e., min(AIC) model) has the smallest expected loss of precision relative to the true, but unknown, model.

It should be noted that AIC does not depend directly on sample size. Bozdogan (1987) noted that, because of this, AIC lacks certain properties of asymptotic consistency and he proposed a related measure, CAIC, by applying his own heuristic to the development of the estimate for Kullback-Leibler information. In particular, for a sample of N cases, $CAIC = -2LL + (\ln(N)+1)p$.

This measure is very similar to the BIC measure proposed by Schwarz (1978) that takes the form $BIC = -2LL + \ln(N)p$, although Schwarz developed his measure as an estimate for a particular posterior Bayes factor not directly related to Kullback-Leibler information. In any case, both CAIC and BIC reflect sample size and have properties of asymptotic consistency although the importance of this property for the interpretation of data for any specific sample setting can be disputed since, unlike significance tests, the interpretation of AIC does not depend on long-range sampling notions. AIC, CAIC and BIC may each be viewed as a penalized log-likelihood (Sclove, 1987) with penalties per parameter of 2, ln(N)+1 and ln(N), respectively. For all reasonable sample sizes, CAIC and BIC apply larger penalties than AIC and, thus, other factors being equal, they tend to select simpler models than does AIC.

Among the reasons for preferring the use of a model selection procedure such as AIC in comparison to traditional significance tests are:

(a) A single, holistic decision can be made concerning the model that is best supported by the data in contrast to what is usually a series of possibly conflicting significance test. Moreover, models can be ranked from best to worst supported by the data, thus, extending the possibilities of interpretation.

(b) Models with various parameterizations can be compared even when the models do not obey hierarchic relations.

(c) Homogeneous and heterogeneous versions of models can be compared; in particular, the homogeneity of variance (homoscedasticity) assumptions required by many significance tests can be circumvented and the selection of the most appropriate model can be based on the information criteria.

(d) Considerations related to underlying distributions for random variables can be incorporated into the decision-making process rather than being treated as an assumption whose robustness must be considered (e.g., models based on normal densities and on log-normal densities can be compared).

Various arguments have been presented against the use of information criteria such as AIC although some of these are difficult to follow. For example, McDonald and Marsh (1990) seem to argue as follows: major premise – the saturated model is always the true model; minor premise – for sufficiently large sample size, AIC will always select the saturated model; conclusion – AIC is defective and cannot be used in practice. In a context such as paired-comparisons among K means, a saturated model based on normal densities would comprise K unique means and variances. Thus, no other model could possibly fit the data better in an absolute sense (i.e., yield a larger log-likelihood). However, if two of the group means are truly equally and very large samples are involved, measures such as AIC will tend to select the correct model, not the saturated model.

As noted above, others are concerned with the fact that AIC does not directly depend upon sample size and, therefore, lacks properties of asymptotic consistency (Bozdogan, 1987). However, variations on AIC such as Schwarz's (1978) BIC and Bozdogan's (1987) CAIC do reflect sample size considerations. In practice, it is not necessarily the case that the property of asymptotic consistency leads to a better procedure in a true-model identification sense.

For example, in the context of comparing non-nested latent clas (mixture) models, Lin and Dayton (1997) found that AIC was superior to BIC when the "true model" was relatively complex (i.e., was based on a relatively large number of parameters). Similarly, Huang and Dayton (1995) report that, for multiple comparisons among bivariate mean vectors, AIC tended to outperform BIC and CAIC when "the null case was excluded and, in general, for heterogeneous cases." However, for multiple regression analysis, the results for AIC and BIC reported by Gagné and Dayton (2002) are more complex but consistent with the observation that AIC is more successful with more complex models.

Clearly, further research around the issue of competing information measures is needed but that does not alter the fact that this class of procedures often provides a highly desirable alternative to traditional significance testing techniques. Finally, it should be pointed out that information measures themselves depend upon certain asymptotic properties of chi-square statistics and, thus, issues of robustness must be considered. This is a researchable topic about which little is known at present. Of course, very similar distributional issues must be considered for significance tests and, despite years of research, the best advice has always been to use large samples.

A technical point about the calculation of AIC (or CAIC or BIC) is that the log-likelihood, LL, often involves the estimation of theoretical variances. The maximum-likelihood estimate for a variance is biased since the denominator for the computation is the sample size, N, regardless of the number of parameters that are estimated in fitting the model to data. In regression analysis with p independent variables, for example, the unbiased estimate for the residual variance is computed by dividing the residual sum of squares by $N - p - 1$ but in the context of computing AIC the divisor for the maximum likelihood estimate is N.

The computation of AIC for any specific model requires the specification of a distributional form (e.g., univariate normal, multivariate normal, multinomial, Poisson, etc.). Then, the log-likelihood, LL, for the sample is computed based on the model and the specified distributional form. In multiple regression analysis, for example, residuals may be assumed to follow a univariate normal density with variances that are homogeneous conditional on the independent variables.

However, unlike conventional significance tests, the set of alternate models being considered may include different specifications and different distributional assumptions. For example, residuals may be characterized as heterogeneous or dependent on the independent variables in various ways. On the other hand, residuals may be assumed to follow a mixture of homogeneous univariate normal densities. In any case, the min(AIC) criterion can be used to order and select among these models.

To illustrate these ideas in the context of real data, consider the plot (Figure 1) for mathematics achievement scores as a function of weekend television watching activity based on a 5% random sample of cases from the public use for the National Education Longitudinal Study (NELS). The distinct non-linear trend based on 1092 cases seems to invite a quadratic regression model (the television watching categories were coded at their upper values except that the final category was coded 6). Conventional F tests for increments to explained variability (ΔR^2) using a direct notation are $F_{linear} = 5.34$, $F_{quad} = 41.05$ and $F_{cubic} = 1.30$. The linear and quadratic terms are significant at the conventional 5% level whereas the cubic term is non-significant. Thus, the three significance tests can be interpreted as supporting the selection of a quadratic model for the data. As reported in Gagné and Dayton (2002), the log-likelihood for homogeneous multiple regression models can be computed directly from the residual sum of squares (SS_e) and sample size:

$$\text{LL} = -.5N \cdot \left[\ln(2\pi) + \ln\left(\frac{SS_e}{N}\right) + 1 \right]. \quad (1)$$

The AIC values for linear, quadratic and cubic models are, respectively, 8140.02, 8101.62 and 8127.30 leading to the choice of the quadratic model as the best approximating model among these three models (using BIC leads to the same preferred model). But, other models might be explored for these data. For example, using the reciprocal of weekend television watching as a predictor (actually, reciprocal of X+1 due to the presence of 0's), the AIC value is 8144.16 which is less preferred than any of the polynomial models. Note that from a conventional point of view, a test of significance can be run for the regression coefficient in the reciprocal regression model ($t = -1.095$, $p = .274$) but there is no direct way of testing the difference in fit between, say, the linear model and the reciprocal model since they are not nested.

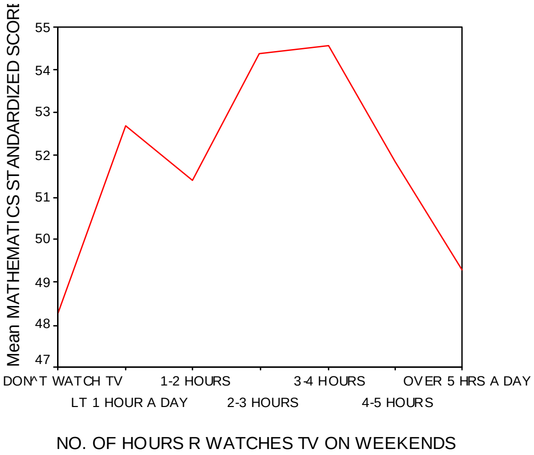

Figure 1. Mathematics achievement scores as a function of weekend television watching.

Paired Comparisons Information Criterion

Dayton (1998, 2003a) proposed a method for comparisons among means using information criteria such as Akaike's AIC. He advocated this approach rather than standard pairwise-comparison procedures such as Tukey tests in order to avoid or minimize the following problems with conventional procedures.

(a) Tukey tests (and variations) have been proposed based on some arbitrary method for controlling the family-wise type I error rate for the set of correlated pairwise contrasts. Release 11.5 of SPSS, for example, provides options for 18 different *post hoc* pairwise comparison procedures that are based on several different approaches to controlling type I error.

(b) Unequal sample sizes and heterogeneity of variance pose difficulties for many procedures. The classic Tukey test, for example, assumes constant sample size and homogeneous variances, an often unrealistic set of assumptions. Modifications of Tukey tests such as Games-Howell tests allow for both unequal sample sizes and heterogeneous variances but only provide approximate control of the family-wise type I error rates by means of an adjustment to degrees of freedom.

(c) Intransitive decisions are routinely encountered with pairwise-comparison procedures in general and pose serious interpretive problems if some overall conclusion is desired for the set of means. For three means in rank order, an intransitive decision entails rejecting the difference between the highest and lowest mean but retaining the null hypotheses for comparisons of these means with the middle mean. It has been argued that this really doesn't pose a problem if the main concern of a study is to draw conclusions about the separate pairwise differences. However, if the focus is on individual pairwise contrasts, what rationale is there for sacrificing power and adopting a family-wise error rate rather than simply running separate t tests for each pair of means?

The method based on information criteria described below and known as paired-comparisons information-criterion, or PCIC, has been the topic of simulations by Cribbie & Keselman (2003) who suggest that PCIC has all-pairs power that is typically superior to standard pairwise comparison procedures (e.g., Tukey

HSD). The method has been extended to repeated observations as well as to data in the form of proportions.

(A) Independent Samples of Means

Consider a design comprising J independent, random groups of respondents with sample sizes, n_j, sample means $\overline{Y}_j$ and unbiased variance estimates, S_j^2, with $N = \sum_{j=1}^{J} n_j$. In PCIC, AIC (or similar measure) is computed for each possible, different ordered subset of means. Thus, only non-overlapping subsets of means are compared rather than all possible subsets. In general, for J groups there are 2^{J-1} distinct patterns of subsets based on ordered means. For example, with three groups with the means ranked and labeled 1, 2, 3, the $2^2 = 4$ ordered subsets are {123}, {1,23}, {12,3}, and {1,2,3,} where a comma is used to separate subsets with unequal means. Focusing on ordered subsets of means and using a min(AIC) [or min(BIC)] strategy avoids the intransitivity problem that may arise when using traditional paired-comparisons techniques without sacrificing interpretability of results.

Assuming homogeneity of variance, the log-likelihood for the m^th model can be written as:

$$LL_m = -\frac{N}{2} Ln(2\pi) - \frac{N}{2} Ln(\hat{\sigma}_W^2) \\ -\frac{1}{2\hat{\sigma}_W^2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \hat{\mu}_{mj})^2 \qquad (2)$$

where $\hat{\sigma}_W^2$ is computed from the ANOVA within-groups sum of squares but with denominator N rather than $N - J$. Means for the m^th model are estimated assuming that the model is correct. The independent parameters estimated for a model comprise the variance and means, as necessary. If variances are assumed to be equal in the same pattern as means, the case is termed the restricted heterogeneous variance case (for other cases, see Dayton, 1998). Assuming the restricted heterogeneous variance case, an estimated variance for a subset of means can be obtained either by pooling variance estimates as appropriate from the separate groups or by computing the (biased) sample variance from the appropriate combined group. For the latter preferred case, the sample variance for a {23} subset of means, for example, would be

$$\hat{\sigma}_{23}^2 = \frac{\sum_{i=1}^{n_2} (Y_{i2} - \hat{\mu}_{23})^2 + \sum_{i=1}^{n_3} (Y_{i3} - \hat{\mu}_{23})^2}{(n_2 + n_3)}. \qquad (3)$$

Assume that, for the m^th model, the pattern of sample means has been partitioned into K non-overlapping subsets. Then,

$$LL_m = -\frac{N}{2}\left[Ln(2\pi) + 1\right] \\ -\frac{1}{2}\sum_{k=1}^{K} n_{mk} \ln(\hat{\sigma}_{mk}^2) \qquad (4)$$

where $\hat{\sigma}_{mk}^2$ is the (biased) variance estimate and n_{mk} is the sample size for the k^th subset.

Table 1 (following page) summarizes NELS data for standardized reading scores for five racial/ethnic as identified in the data base. Tukey tests, as well as Games-Howell tests that lack the homogeneity of variance assumption, yield a typical intransitive pattern of differences with three overlapping, non-significant ranges comprising, in rank order of means from high to low, {123}, {34} and {45}. The three smallest AIC values assuming homogeneity of variance and not making this assumption are shown in Table 1.

Note that min(AIC) occurs for the pattern {12,345} assuming the restricted heterogeneous variance case although several models show quite similar AIC values. An interesting feature of model comparisons with AIC and related information measures is that, although a single preferred model is identified, a ranking of alternative models is provided. Additional illustrative analyses for both the homogeneous and heterogeneous cases are presented in Dayton (1998, 2003a) as well as in connection with a Gauss program (Aptech Systems, 1997) for conducting these tests (Dayton, 2001a).

Table 1
NELS Reading Standardized Scores

				Homogeneity		Restricted Heterogeneity	
"Race"	n	Mean	Variance	Pattern	AIC	Pattern	AIC
White Non-Hispanic	798	52.55	98.21	{1,2,345}	8926.90	{12,345}	8926.62
API	75	50.40	97.66	{1,2,3,45}	8927.59	{1,2,345}	8927.37
Hispanic	140	47.36	92.13	{12,34,5}	8928.30	{12,3,45}	8927.56
Black Non-Hispanic	152	46.16	77.68				
American Indian	44	46.00	70.39				
	1209						

(B) Means Based on Repeated Observations

Consider a cohort of individuals that is measured on the same variable at several points in time. Assuming multivariate normality, the parameters of the distribution are means and variances for the occasions of measurement as well as covariances among occasions. As with independent observations, attention is focused on the distinct ordered subsets of means and AIC, or a related information measure, can be used to select a preferred pattern.

As for the case of independent groups, variances and covariances can be homogeneous, heterogeneous or restricted heterogeneous. However, the situation is more complex since these conditions can be applied separately to the variances, covariances or to both. In addition, various patterned covariance matrices may be considered to be appropriate (e.g., a simplex pattern with observations closer in time more highly correlated that those further apart in time). Dayton (2003a) presents more detailed information about this case along with illustrative data.

(C) Independent Samples of Proportions

Consider J groups of sizes n_j with sample proportions, $p_1, p_2, \ldots, p_J$, for a dichotomous dependent variable. The theoretical model for the data is that responses represent a series of 0/1 Bernoulli trials with a true population probability, π_j, of a favorable outcome (e.g., 1 or positive) for the jth group. The log-likelihood for any specific ordered outcome (e.g., 0110 for proportions based on four outcomes) in the jth group is $n_j p_j \ln(p_j) + n_j(1-p_j)\ln(1-p_j)$ and the log-likelihood for the total sample is found by summing across the J groups:

$$LL = \sum_{j=1}^{J} \left[n_j p_j \ln(p_j) + n_j(1-p_j)\ln(1-p_j) \right]. \tag{5}$$

Note that $n_j p_j$ is the expected number of favorable outcomes and $n_j(1-p_j)$ is the expected number of unfavorable outcomes. The sample proportion, p_j, is the MLE for the corresponding population proportion. Unlike the situation for sample means, there is no need to consider homogeneous and heterogeneous cases since each Bernoulli process is based on a single parameter, π_j. Otherwise, model selection follows the same reasoning as for independent sample means (Dayton, 2001a). That is, there is a total of 2^{J-1} distinct patterns of subsets of proportions to evaluate and proportions for a model are estimated assuming that the model is correct. Illustrative analyses for this case are presented in Dayton (2001a, 2003a).

PCIC for Distributions

Standard pairwise comparison procedures, such as Tukey HSD and its many variations, have been the subject of a good deal of research directed toward assessing their robustness with respect to distributional assumptions. Typically, non-normal distributions with varying degrees of skew and kurtosis are selected for comparison (e.g., Keselman, Lix & Kowalchuk, 1998, report simulations with normal distributions, three-degree-of-freedom chi-square distributions and a highly non-normal distribution with skewness and kurtosis indices equal to 6.2 and 114, respectively). The issue, then, is the degree of sensitivity of the multiple comparison procedures to departures from normality. Also, a number of simulations have dealt with the relative power of pairwise comparison procedures (e.g., Ramsey, 2002).

An alternative approach is to directly model the underlying distributions for observed data and then compute appropriate likelihoods for candidate distributions of interest. Once these distributions have been selected, procedures comparable to PCIC can be implemented. In practice, identifying the set of candidate distributions is a non-trivial problem. Two classes of plausible models that have credibility in practice than can be compared are normal and log-normal densities.

The motivation for log-normal models arises from the fact that, in contrast to an additive effect, a multiple effect for an independent variable can be modeled in log-linear terms. For example, the usual additive model for a response in a one-way ANOVA design can be represented as $Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$ where μ is a grand mean effect, τ_j is the effect of the jth treatment and ε_{ij} is a residual error term. Alternatively, assuming a multiplicative, rather than an additive treatment effect, yields the model: $Y_{ij} = \mu \times \tau_j \times \varepsilon_{ij}$ or $\ln(Y_{ij}) = \mu^* + \tau_j^* + \varepsilon_{ij}^*$ where the * superscript denotes a parameter on a logarithmic scale. In practice, many positively skewed distributions of observations are reasonably well approximated by log-linear models.

Some preliminary simulation results have been carried out for two-sample and a limited number of three-sample cases to assess how well the AIC and BIC information measures distinguish between samples based on normal and log-normal distributions (Dayton, 2003b). In one series of simulations, theoretical log-normal densities with means, standard deviations of (0, .1), (0, .5) and (0, 1.0) in log units corresponding to (1.00, .10), (1.13, .60) and (1.65, 2.16) in raw units were considered. The first distribution is slightly skewed (index = .30) and modestly kurtotic (index = 3.16), the second distribution is moderately skewed (index = 1.75) and somewhat peaked (index = 8.89), while the third distribution is both highly skewed (index = 6.18) and highly kurtotic (index = 113.94). In a second series of simulations, information criteria were compared assuming only log-normal densities but the generated data were either normal or log-normal.

Typical results for two groups are, in additional to the expected sample size differences: (a) BIC selected the correct model more often than AIC in virtually all simulated cases with an average difference ranging from about 6% to 13%; (b) both information criteria were much more successful in selecting models when the true distribution was log-normal as opposed to when it was normal. This latter result occurs because, as the median increases, log-normal distributions assume a nearly symmetric shape that approximates normality. Limited results for three samples suggest that, as was true for two groups, BIC tends to select the correct pattern of means more often than does AIC and both criteria were more successful for log-normal than for normal distributions. The superiority of BIC over AIC should not be generalized at this time, however, since Dayton (1998) found for cases with several groups that neither criterion was uniformly superior to the other.

Number of Components in Mixture models

An emerging area of interest in applied research is the use of finite mixture models when distributions such as normal, Poisson and binomial fail to provide satisfactory fit to data. An impetus for considering mixtures is the phenomenon of over-dispersion which is manifested by, for example, distributions with "heavy tails." For situations of this sort it is often reasonable to assume that observations represent a mixture from two or more sub-populations rather than arising from a single population. In general, a mixture of J distributions for some dependent variable, Y, can be represented by: $g(Y|\beta) = \sum_{j=1}^{J} \theta_j \times g_j(Y|\beta_j)$ where θ_j are mixing fractions such that $\sum_{j=1}^{J} \theta_j = 1$, g( ) is some specified probability (e.g., binomial) or density (e.g., normal) function based on a vector of parameters, β_j, and β is a vector containing all relevant parameters.

For a mixture of two heterogeneous normal densities, for example, $g_1(Y \mid \mu_1, \sigma_1^2)$ and $g_2(Y \mid \mu_2, \sigma_2^2)$ would represent normal densities with unique means and variances that are mixed in proportions θ_1 and $\theta_2 = 1 - \theta_1$. To fit such models to data, the parameters for the separate components as well as mixing fractions must be estimated. For a mixture of two normal densities this would entail estimating five unique parameters (two means, two variances and one mixing proportion). Some relatively simple mixtures (e.g., normal densities) can be estimated using available statistical software such as Mplus (Muthén and Muthén, 1998) but specialized programs such as LEM (Vermunt, 1993) are required in more complex cases such as latent class models.

A persistent dilemma for applications of mixture models is that models with varying numbers of components cannot be compared using conventional significance tests even though these models are hierarchical. For example, the comparison of a mixture of two normal densities to a single normal density could, seemingly, be based on a difference-chi-square test since the single normal density is a restricted form of the mixture (e.g., by setting $\theta_2 = 0$). However, as noted by Everitt and Hand (1981) and Titterington, Smith and Makov (1985), among others, this difference-chi-square statistic fails to satisfy theoretical requirements related to boundaries of the parameter space and is not distributed as expected (nor is its distribution known). Some insight into the problem can be seen from observing that the single restriction, $\theta_2 = 0$ is equivalent to the two restrictions $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ since, in either case, the resulting model is a single normal density. In fact, the mixture is based on five parameters whereas the single normal distribution is based on only two parameters, yet only one restriction is required to obtain the simpler from the more complex model.

Given the failure of conventional significance tests to provide a basis for assessing the number of components in a mixture, information measures such as AIC present an attractive alternative. Information criteria provide a single summary statistic for each model being compared and avoid the asymptotic distributional issues faced by difference-chi-squares tests for mixture models. Some preliminary work on assessing AIC, BIC and related measures was reported by Dayton (2001b) who focused on the issue of selecting the appropriate number of mixtures in binomial models (restricted latent class models) with four and six binary variables.

Simulations were based on samples sizes ranged from 80 to 1280, binomial probabilities for mixtures of two and three processes were selected to represent varying degrees of discriminability of the components and mixing proportions were varied from equal splits to cases where one component represented only 20% of the cases. Cases with high discriminability involved, for two components, cases with binomial probabilities and .1, .5 and .1, .8 where low discriminability involved cases with binomial probabilities of .1, .2. All of the measures studied provided reasonable correct identification rates for the high disciminability cases (e.g., 80% and above across the conditions) but very poor correct identification rates for the low disciminability cases (e.g., 10% or less across the conditions). Dayton (2001b) concludes that this area of analysis requires "…reasonably large sample sizes and the realization that poorly defined latent structures will almost certainly go undetected."

Conclusion

Although the recommendation has been repeated often in the past, researchers should become aware of modern alternatives to the use of significance tests when comparing alternate models is the focus of analysis. Information theoretical procedures such as Akaike AIC provide a holistic approach to ordering and selecting among competing models that avoids the piece-meal and potentially inconsistent outcomes that arise from applying multiple significance tests. This paper has summarized applications of these measures to multiple comparisons including the possibility of varying distributional assumptions and to mixture models where traditional significance tests are known to be inappropriate.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 30, Part A 9-14.

Aptech Systems, Inc. (1997). GAUSS for Windows NT/95: Version 3.2.32, Maple Valley, WA.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61*, 287-292.

Cribbie, R. A. & Keselman, H. J. (2003). A power comparison of pairwise multiple comparison procedures: A model testing approach versus stepwise procedures. *British Journal of Statistical & Mathematical Psychology*, 56, 157-182.

Dayton, C.M. (1998). Information criteria for the paired-comparisons problem. *American Statistician, 52,* 144-151.

Dayton, C. M. (2001a). SUBSET: Best subsets using information criteria, *Journal of Statistical Software,* Vol 6., Issue 02, April.

Dayton, C. M. (2001b). Performance of information criteria for number of components for product-binomial processes. Paper presented at the Mixtures 2001: Recent Developments in Mixture Modeling conference at Universitat der Bundeswehr, Hamburg, Germany, July.

Dayton, C. M. (2003a). Information criteria for pairwise comparisons. *Psychological Methods, 8*, 61-71.

Dayton, C. M. (2003b). A modeling approach to *post-hoc* comparisons of means and proportions. Paper presented at Second Workshop on Research Methodology (RM2003), Vrije University, Amsterdam, The Netherlands, June.

de Leeuw, J. (1992). Introduction to Akaike (1973) Information theory and an extension of the maximum likelihood principle. In S. Kotz & N. L. Johnson (Eds) *Breakthroughs in Statistics Volume I Foundations and Basic Theory*. New York: Springer-Verlag.

Everitt, B. S. & Hand, D. J. (1981). *Finite Mixture Models*. New York: Chapman & Hall, Ltd.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference* (2nd Edition). New York: Hafner.

Gagné, P. & Dayton, C.M. (2002). Best regression model using information criteria. *Journal of Modern Applied Statistical Methods, 1*, 479-488.

Heermann, E. & Braskamp, L. A. (editors) (1970). *Reading in Statistics for the Behavioral Sciences.* New Jersey: Prentice-Hall.

Huang, C-C & Dayton, C.M. (1995). Detecting patterns of bivariate mean vectors using model-selection criteria. *British Journal of Mathematical & Statistical. Psychology, 48*, 129-147.

Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods, 3*, 123-141.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Lin, T. S. & Dayton, C. M. (1997). Model-selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, *22*, 249-264.

McDonald, R. P. & Marsh, H. W. (1990). Choosing a multivariate model: noncentrality and goodness of fit. *Psychological Bulletin, 107*, 247-255.

Muthén, L. K., and Muthén, B. O. (1998), *Mplus User's Guide*. Los Angeles: Muthén and Muthén.

Neyman, J. & Perarson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society (A), 231*, 289-337.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods, 7*, 504-523.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62*, 157-176.

Titterington, D. M. Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Models*. New York: John Wiley & Sons.

Vermunt, J. K. (1993). Log-linear & event history analysis with missing data using the EM algorithm. WORC Paper, Tilburg University, The Netherlands.