Chapter 9

Assessing Statistical Techniques for Detecting Multispecies Samples of Heteromyids in the Fossil Record: A Test Using Extant *Dipodomys*

MARC A. CARRASCO

ABSTRACT

Sixteen dental measurements in nineteen species of the extant rodent genus *Dipodomys* were examined to determine which techniques commonly used to identify the presence of multiple species in qualitatively homogeneous fossil samples are reliable. Each technique was tested using a simulation approach whereby samples created from a sympatric pooled-species group were compared with those of a single-species referent to determine the power of each technique. The type-I error rate of each method was assessed by comparing simulated pooled samples created from a single species to the same referent. Most techniques, including all range-based methods, performed poorly. Only the coefficient of variation using a 1% significance level and Levene's test of relative variation using a 2.5% significance level were reliable. The most useful dental variables were the widths of the upper and lower first and second molars.

INTRODUCTION

Species recognition in the fossil record is an ongoing problem. Groups that lack qualitative differences among taxa, such as rodents and primates, have forced workers to use quantitative variables to address taxonomic questions (e.g., Barnosky, 1986; Martin and Andrews, 1993; Carrasco, 1998). Many observed distributions of closely related sympatric species frequently overlap, thus obscuring taxonomic boundaries and masking the true taxonomic diversity of a fossil assemblage (Plavcan, 1993). To this end, more than ten different techniques have been proposed, using only quantitative characters (primarily dental characters), to recognize the presence of closely related sympatric species in fossil samples. Despite several attempts to compare select groups of these methods (e.g., Cope and Lacy, 1992, 1995; Martin and Andrews, 1993; Donnelly and Kramer, 1999), considerable controversy exists over which technique is the best. In addition, previous studies on the efficacy of these methods have primarily used primates as the study group (e.g., Gingerich, 1974; Cope and Lacy, 1992; Cope, 1993; Martin and Andrews, 1993; Plavcan, 1993; Cameron, 1998). Thus the generality of those conclusions to other taxonomic groups is unclear.

In an attempt to determine how broadly applicable the previous studies' results are, this study investigates which, if any, of the most commonly employed species-recognition methods are useful in a nonprimate group, the extant heteromyid rodent *Dipodomys*. *Dipodomys* was selected because fossil samples of heteromyids are composed primarily of isolated teeth with a simple, conservative dental morphology (Wahlert, 1993). This morphology has forced workers to rely on quantitative characters to diagnose fossil heteromyid species (e.g., Barnosky, 1986; Wahlert, 1993; Carrasco, 1998). Therefore, results obtained here will be directly applicable to problematic fossil samples. To test the various procedures with extant taxa, a simulation approach was used whereby ''simulated'' samples of defined siz-

es were created from both a multispecies sample of closely related sympatric taxa and a large single-species sample. These simulated samples were then compared to closely related single-species referents to determine the reliability of each method.

MATERIALS AND METHODS

SAMPLES

More than 3700 specimens from 19 different species were analyzed. See Carrasco (1999) for a complete list of specimens examined and locality information. Each species was separated into geographically restricted pooled samples, with the assumption that a smaller range of geographic sampling correlates with less intraspecific variation. In total, 50 geographically restricted pooled samples were created with sample sizes ranging from 22 to 132. All pooled geographic samples encompassed a range with a radius less than 225 kilometers.

A multispecies pooled sample was created by pooling available specimens of three sympatric species from a limited geographic region (less than a 110 km radius). The multispecies sample was from northern Baja California and included a total of 274 specimens of *D. gravipes*, *D. merriami*, and *D. simulans* (see appendix 9.1). The mean body sizes of these three sympatric species are significantly different from each other, providing a method to detect multiple species. Despite these significant differences in mean, the observed range of each measurement overlaps considerably for each species pair. This study attempts to answer the question so frequently unknown in paleontological samples—if a random sample were collected from this region, would it be possible to determine whether several species were present based solely on an analysis of tooth size?

MEASUREMENTS

The dental measurements were those most commonly taken by paleontologists—maximum lengths and widths of teeth. In all, sixteen dental variables were measured on each specimen: length and width of upper premolars (APP4 and TP4, respectively), upper first molars (APM1, TM1), upper second molars (APM2, TM2), upper third molars (APM3, TM3), lower premolars (app4, tp4), lower first molars (apm1, tm1), lower second molars (apm2, tm2), and lower third molars (apm3, tm3). All dental measurements were taken through an Ehrenreich Photo-Optical Industries ''Shopscope'' with 10X magnification and a precision digital positioner. Measurements were taken to the nearest 0.01 mm and entered directly into the computer program Statistica 5.1h. One hundred specimens were remeasured to test the accuracy of the measurement technique. All variables displayed an average relative error (the absolute error of a measurement divided by its original measurement multiplied by 100) of less than 2%.

Specimens were placed into one of six age groups (juvenile, 1, 2, 3, 4, or 5) according to the criteria used in Carrasco (2000a). Because previous work on the teeth of *Dipodomys* uncovered a significant degree of age variation (Carrasco, 2000b), juvenile and age-5 specimens were eliminated from this study. Despite a significant degree of age variation in the remaining four age groups, all four were maintained to reflect the age variation typically seen in fossil samples.

THE SIMULATION PROCEDURE

1. Reference standards (the samples to which all others will be compared; see below for more information) for each technique and each variable were selected from a set of 50 geographically restricted single-species pooled samples.
2. Using a random number generator, 100 single-species samples at each of four sample sizes (n = 5, 10, 20, and 35) were drawn from the geographically restricted single-species pooled sample with the largest sample size (*D. agilis*). Each sample was drawn without replacement.
3. The three-species sympatric pooled sample was divided into three groups, each of which contained a different two-species pair: *D. gravipes*/*D. merriami*, *D. gravipes*/*D. simulans*, and *D. merriami*/*D. simulans*. Fifty multiple-species samples were drawn at random from each of the two-species paired pools as well as from the original three-species sympatric pooled sample. This procedure was repeated at four sample sizes (n = 5, 10, 20, and 35) with each sample drawn without replacement. No attempt was made

to equilibrate the different probabilities of selecting one species over another caused by species sample size differences (see appendix 9.1). By not correcting for these specimen number differences, a wider range of relative species' percentages in the simulated samples was created. Tests for the presence of multiple species in these simulated pooled-species samples are therefore testing groups with varying sample sizes of each species as might be encountered in a fossil sample.

4. Values for each of the techniques for each variable were calculated for all simulated single-species and pooled-species samples.
5. Values of the reference standards for each technique were then compared to the values of the 400 simulated single-species samples and the 800 pooled-species samples to assess the type-I error rate and the power of each technique.

Reference Standards

As suggested by Cope and Lacy (1995), a single reference standard was selected from a variety of samples that are as closely related and/or ecologically similar to the simulated fossil samples as possible. In this study, the reference set of species was composed of the 50 geographically restricted pooled samples of *Dipodomys*. Three methods for selecting a single reference standard (referent) from these 50 samples were tested: a reference standard composed of the maximum values for each variable (e.g., Martin and Andrews, 1993; Cameron, 1998), a referent composed of the median values for each variable (Cope and Lacy, 1992, 1995), and a referent using only the values from the sample with the largest sample size (largest-n referent; Cope and Lacy, 1992, 1995). Because the results across all techniques using the maximum-value referent were markedly worse than those of the other referents, and because the median and largest-n referents performed similarly, only the results using the median referent will be reported here. For a detailed description of the results obtained using the maximum and largest-n referents see Carrasco (1999).

Determination of Type-I Error Rate and Power

The type-I error rate, the rate at which the null hypothesis is incorrectly rejected, was estimated by comparing each simulated single-species sample to each of the reference standards. A type-I error occurred when the statistic for a given variable in a simulated sample exceeded that of a reference standard. The type-I error rate is the total number of type-I errors divided by the total number of comparisons made. Any variable of a particular technique with a type-I error rate greater than five percent is deemed to have a high type-I error rate.

Power, the probability of rejecting the null hypothesis when it is not true, was tested by comparing each of the 800 simulated pooled-species samples to each reference standard. The power of a technique for a given variable is the number of times the null hypothesis was rejected divided by the total number of comparisons made multiplied by 100. The power ranges from 0 to 100 with higher numbers indicating a more powerful method.

Techniques

Univariate methods are by far the most commonly used techniques to evaluate a multiple-species hypothesis (e.g., Gingerich, 1974; Cope and Lacy, 1992, 1995; Cope, 1993; Martin and Andrews, 1993; Plavcan, 1993; Fuller, 1996; Cameron, 1998). While multivariate techniques have the potential to offer a different, and perhaps improved, approach toward identifying the taxonomic composition of a fossil assemblage, in practice, fragmentary fossils and small sample sizes often limit the utility of multivariate methods (Plavcan, 1993; Cope and Lacy, 1995). In this paper, only tests that were applicable to such paleontological samples and also could be readily used by investigators not schooled in advanced statistical methods were evaluated. Thus analyses, such as the Fligner and Killeen method advocated by Donnelly and Kramer (1999), which require more complex statistical programming, were not evaluated here. For these reasons, this paper tests the coefficient of variation (CV), max/min index (MI), range as a percentage of the mean (R%), and Levene's test of relative variation (LEV) methods.

The CV is defined as $100*SD/\bar{x}$ where SD is the standard deviation of the sample and $\bar{x}$ is the sample mean. A variety of statistical

methods have been proposed to compare CVs (e.g., Lande, 1977; Sokal and Braumann, 1980). While most of these tests lack power at small sample sizes (Cope and Lacy, 1992), the method advocated by Sokal and Braumann (1980) appears to be a reliable method to assess the taxonomic composition of a fossil assemblage (Cope and Lacy, 1992). To compare CVs using this procedure, the standard error of the CV is calculated according to the formula,

$$\sqrt{\left(\frac{V^2}{2n}\right)\left(\frac{n}{n-1}+2V\right)\left(1+\frac{1}{4n}\right)^2}$$

where V is the CV/100 of the referent and n is the sample size of the fossil assemblage (Sokal and Braumann, 1980 as suggested by Cope and Lacy, 1992). The sample-size bias correction factor (V*) used by Sokal and Braumann (1980) was avoided following the recommendations of Cope and Lacy (1992) and Cope (1993). The standard error of the CV is then used to create a one-sided confidence interval for the reference sample that is compared to the CV of the fossil sample. A second approach tested here is the ''simulation approach'' of Cope and Lacy (1992, 1995) and Cope (1993). This approach involves Monte Carlo simulations in which a Pascal computer program (Cope and Lacy, 1992) creates a 95% confidence interval for each CV of the reference group. The CV of the fossil sample is then compared to the corresponding confidence intervals. One-tailed tests were employed and evaluated at three significance levels (5%, 2.5%, and 1%) for the formula-based method of Sokal and Braumann (1980) and two significance levels (5% and 1%) for the simulation approach.

Two different range-based methods were evaluated: MI and R%. MI compares the ratio of the maximum value to the minimum value of the referent pooled sample for a given variable to the same ratio of the fossil sample. If the ratio of the fossil sample exceeds that of the referent, multiple species are hypothesized to be present. Similarly, the null hypothesis is rejected when R%, defined as $100*OR/\bar{x}$ (where OR is the observed range of the sample and $\bar{x}$ is the sample mean), of the fossil sample exceeds that of the referent. The referent used for these two methods has typically been composed of the maximum values of a set of reference species (Martin and Andrews, 1993; Cameron, 1998). However, as noted above, all methods tested that used a maximum referent performed poorly. An alternative approach suggested by Cope and Lacy (1995) is to adopt a ''simulation approach'' with MI and R%, identical to their CV simulation approach. As was done with the CV, the results of the simulation approach for MI and R% were evaluated using 5% and 1% significance levels.

Following Schultz (1985), the LEV procedure consists of transforming each value, X_i, of a sample of values to Y_i according to the formula, $|X_i - medX|/medX$ where $medX$ is the median of the set of sample values. An ANOVA is then used to determine whether there are significant differences between the group means of the extant referent sample and fossil sample of transformed variates, Y_i, for a given variable. The greater the variance in the sample the greater the group mean will be. Comparisons were made at two one-tailed significance levels, 5% and 2.5%.

Methodological Assumptions

Each technique is designed to use quantitative data to detect multiple-species samples by comparing the morphological variation of the fossil sample with the variation in single-species extant taxa. The underlying assumption of this method is that fossil taxa are no more variable than living taxa. While this uniformitarian approach has been criticized by some authors (Kelley, 1993; Kieser, 1994), it has been shown by numerous others that there appears to be a consistent pattern of metric dental variation across all mammalian species (e.g., Simpson and Roe, 1939; Gingerich, 1974; Yablokov, 1974; Martin and Andrews, 1993).

Implicit in this assumption is that the sources of variation in both the fossil and extant groups are similar. Temporal, geographical, secondary sexual, and age variation can all significantly affect the total morphological variation of a group. Because it is virtually impossible to assign a sex and sometimes an age to individuals in a fossil sample in which there might be multiple species, extant reference groups should contain

both sexes and encompass most age groups so that the variation effect of these factors is comparable to that of the fossil sample. To limit the effects of geographic variation, a clear understanding of the geographic range from which the fossil and extant samples are collected should be obtained. While most single fossil assemblages contain individuals from a relatively limited geographic region (Martin and Andrews, 1993), it is possible that a single sample could have been collected from a wide catchment area. In particular, fossil groups that consist of multiple samples from different localities should be tested with caution. If this is the case, it might be best to compare the fossil group to an extant reference group from a large geographic range (Albrecht and Miller, 1993; Martin and Andrews, 1993). At the same time, because of the generally limited geographic extent of fossil assemblages, extant comparison groups should be from a limited geographic range to increase the power of the technique. Unfortunately, temporal variation is difficult to equilibrate between fossil and living groups. Although Martin and Andrews (1993) assert that time-averaging needs to be demonstrated, rather than assumed, every attempt should be made to identify and limit the temporal range of the fossil sample.

In addition, all of the techniques presented here can be used to test only a single-species null hypothesis. Previous studies have shown that the morphological variation in a multiple-species sample is often indistinguishable from that of a single-species sample (Cope and Lacy, 1992, 1995; Cope, 1993; Plavcan, 1993). Therefore, the presence of a single-species sample can never be conclusively demonstrated. In addition, falsification of the single-species null hypothesis does not necessitate that a fossil sample is composed of more than one species (Donnelly and Kramer, 1999). Falsification indicates only that the sample has an abnormally high level of variation, which might be accounted for by several causes such as those outlined above. Therefore, it is necessary to minimize the possibility of such confounding factors.

RESULTS

The average type-I error rate and power results across all samples sizes tested are displayed in Tables 9.1 and 9.2. Because these average values reflect the overall pattern found in each technique, results will not be discussed by sample size. Sample-size effects will be addressed in a later publication.

Type-I Error Rate

The CV formula type-I error rates were generally low except those that used a one-tailed 5% significance level. This 5% significance method had average type-I error rates greater than 5 in nine variables. The 1% formula method had a low overall average type-I error rate (2.5), while the 2.5% method rate averaged about 5. The average type-I error rate for the 5% CV program method was consistently higher than any other CV method tested—7 or greater across the majority of variables. The 1% program method had lower average values, which were slightly greater than those of the 1% CV formula method. Across all CV methods, the lowest rates were seen in the upper and lower premolars, TM1, TM2, tm3, and apm1.

The average MI type-I error rates of the 5% method frequently exceeded 5, while those of the 1% technique generally had average rates of less than 2. As seen in the CV methods, upper and lower premolar dimensions, TM1, TM2, tm3, and apm1 had low type-I error rates. Similar to the MI methods, the 5% R% method had high average type-I error rates (average rate of 10.1), while the 1% method had generally low average rates of 3 or less. The variables with the lowest type-I error rates were the lower premolar dimensions, APM1, TM1, apm1, and tm3.

The overall average rate in the 5% LEV method slightly exceeded the 5.0 critical level whereas the 2.5% method average was less than 3.0. All premolar dimensions, TM1, TM2, apm2, and tm2 had low type-I error rates.

Power of the Techniques

Both the 2.5% and 5% CV formula methods had an overall average power greater than 30. The 1% method had an average power of 23.8. The 5% program method was the most powerful CV technique, with 10 variables having an average power greater than 40. The 1% program method had a pow-

TABLE 9.1

Average Type-1 Error Rate of Each Method by Variable

See Methods section for dental abbreviations. CVF1, CVF2.5, and CVF5, coefficient of variation formula methods using 1%, 2.5%, and 5% significance levels, respectively; CVP1 and CVP5, coefficient of variation program methods using a 1% and 5% significance level, respectively; MI1 and MI5, max/min index methods using a 1% and 5% significance level, respectively; R%1 and R%5, range as a percentage of the mean methods using a 1% and 5% significance level, respectively; LEV2.5 and LEV5, Levene's test of relative variation methods using a 2.5% and 5% significance level, respectively.

	Method										
Variable	CVF1	CVF2.5	CVF5	CVP1	CVP5	MI1	MI5	R%1	R%5	LEV2.5	LEV5
APP4	0	1	1	0	3	0	0	1	7	1	2
TP4	0	1	1	1	4	0	2	1	3	1	1
APM1	2	6	9	2	12	0	2	0	3	5	9
TM1	0	0	0	0	1	0	0	0	0	0	0
APM2	13	18	27	15	30	5	15	17	35	2	7
TM2	0	0	1	0	2	0	1	0	7	0	0
APM3	6	12	22	8	25	3	23	6	24	2	4
TM3	10	14	21	11	24	1	4	3	18	3	7
app4	0	1	2	0	3	0	0	0	2	1	3
tp4	0	1	3	1	5	0	1	0	3	1	2
apm1	0	2	6	1	7	0	2	0	2	3	7
tm1	5	9	14	6	19	12	20	12	23	7	13
apm2	2	6	10	3	11	7	31	1	3	2	3
tm2	1	4	9	2	11	1	6	13	25	1	2
apm3	1	4	7	2	7	1	9	1	9	1	3
tm3	0	1	2	0	5	0	0	0	2	3	6
AVG.	2.5	5.1	9.0	3.4	11.4	1.6	6.7	3.2	10.1	2.7	5.3

TABLE 9.2

Average Power of Each Method by Variable

See Methods section for dental abbreviations and table 9.1 for statistical abbreviations.

	Method										
Variable	CVF1	CVF2.5	CVF5	CVP1	CVP5	MI1	MI5	R%1	R%5	LEV2.5	LEV5
APP4	21	28	35	24	38	2	13	35	52	10	16
TP4	23	30	37	26	41	7	18	10	25	19	26
APM1	26	33	41	30	43	2	18	3	21	28	35
TM1	28	35	42	32	45	7	22	0	2	18	26
APM2	27	35	43	30	46	4	20	32	49	17	24
TM2	45	54	62	50	66	31	55	50	65	34	43
APM3	10	16	20	11	24	10	21	11	25	5	8
TM3	30	37	43	34	47	0	7	12	34	33	39
app4	7	12	16	9	19	0	2	1	10	5	9
tp4	5	10	12	6	14	0	2	2	10	4	7
apm1	20	29	36	25	39	3	18	4	17	22	29
tm1	29	36	44	34	49	23	43	28	51	33	41
apm2	23	30	37	26	41	29	51	6	20	16	22
tm2	49	58	66	54	73	20	46	50	71	52	60
apm3	16	23	27	18	29	17	28	15	25	11	17
tm3	22	28	35	25	40	2	11	9	28	28	34
AVG.	23.8	30.8	37.2	27.3	41.0	9.6	22.8	16.8	31.9	21.4	27.7

er average that was comparable to those of the 1% and 2.5% formula techniques. Across all CV techniques, TM2 and tm2 had the highest powers, with average values greater than 40 for all methods.

The 5% MI method had an overall average power of 22.8, and the 1% method averaged 9.6. Slightly more powerful, the 5% R% method averaged 31.9 while the 1% technique had an average power of 16.8. TM2 and tm2 displayed the greatest powers in both the MI and R% methods.

LEV methods exhibited average power comparable to the 5% R% and MI methods with an overall average power just less than 30 in the 5% LEV method and around 20 in the 2.5% method. TM2, TM3, tm1, and tm2 exhibited the most power.

DISCUSSION

Previous work has suggested that the best variables to distinguish closely related sympatric fossil taxa are posterior tooth dimensions, in particular those of the first and second molars (Gingerich, 1974; Cope and Lacy, 1992, 1995; Cope, 1993; Plavcan, 1993). These conclusions are confirmed in this study, with a few notable exceptions. Five of the eight first and second molar dimensions consistently had powers greater than 20 and type-I error rates less than 5—APM1, TM1, TM2, apm1, and tm2. However, APM2 and tm1 were among the variables with the highest type-I error rates, while apm2 had lower powers. In addition, contrary to the findings of previous workers, the two variables TP4 and tm3 had powers and type-I error rates similar to those of the first and second molars. Another item to note is that the width dimension of every tooth exhibited a higher power than the accompanying length dimension while the type-I error rates of the two classes of dimensions were similar. This result is opposite to that found in primates, where lengths were found to be more powerful than widths (Cope, 1993; Cope and Lacy, 1995), but in line with previous work on kangaroo rats that found a greater degree of variability in posterior length dimensions relative to width dimensions (Carrasco, 2000b).

Overall, no statistical method was clearly more useful than all others. Previous workers had found the CV formula and program methods, using a 95% confidence interval, to be quite powerful while committing few type-I errors (Cope and Lacy, 1992, 1995; Cope, 1993). While displaying a considerable degree of power, the type-I error rate of these methods exceeded the stated 0.05 error rate for many variables, especially the posterior tooth dimensions. This type-I error rate contrasts with the low type-I error rate found by Cope (1993) and Cope and Lacy (1992, 1995) in their work on cercopithecine primates, suggesting that the 95% methods do not always meet the stated error rate of the analyses and are therefore not universally appropriate. The results of this study also confirm the suspicions of other workers who have claimed that the use of the CV to assess the taxonomic diversity of a fossil assemblage can result in an unacceptably high number of type-I errors (Martin and Andrews, 1993; Donnelly and Kramer, 1999). This poor performance may be due to comparing distributions that are similar in shape and either normal, leptokurtotic, or strongly skewed (Donnelly and Kramer, 1999). In a cursory examination of the underlying distributions, no significant correlation was found between the shape of the distributions in the CV type-I error rate analyses and the overall CV type-I error rates (this lack of correlation was also found in the R%, MI, and LEV techniques). However, a more thorough analysis of these distributions, which is beyond the scope of this paper, is needed to reach a more definitive conclusion on how the underlying distribution patterns might have affected the results. Nevertheless, while there is a slight sacrifice in power, a more reasonable type-I error rate was recovered in the 99% confidence interval methods.

Results of the range-based program methods (MI and R%) were different from those found in previous studies. Employing the 95% confidence intervals of these programs, Cope and Lacy (1995) found that these techniques (using the largest-n referent) produced an acceptable type-I error rate with a power slightly less than that found with the CV program methods. They also concluded that range-based program methods using a median referent had an unacceptably high type-

I error rate. In this study, the average type-I error rate of the median referent MI techniques was less than that of the largest-n referent. In addition, the empirical type-I error rates of all of the 95% confidence interval program methods exceeded the stated 0.05 rate of the analyses. The techniques that used the 99% confidence interval had much lower type-I error rates, but the lowest powers of all methods tested. Overall, the range-based program methods performed more poorly than the 1% and 2.5% CV formula and program methods.

The results of Levene's test of relative variation were promising. These methods displayed a relatively low type-I error rate while exhibiting average powers greater than 20.0. The 5% method had slightly elevated type-I error rates in four of the eight first and second molar tooth dimensions, limiting its utility. However, the 2.5% LEV average type-I error rate and power were acceptable and comparable to those of the CV 1% program method. The results of this study support the conclusions of Schultz (1985) and Donnelly and Kramer (1999).

CONCLUSIONS AND RECOMMENDATIONS

Utilization of statistical methods to detect the presence of closely related sympatric species from a single morphologically homogeneous fossil assemblage composed solely of teeth is a common procedure in paleoanthropology (e.g., Gingerich, 1974; Cope, 1993; Martin and Andrews, 1993; Plavcan, 1993). The purpose of this study was to determine which, if any, of the most common statistical methods was the most reliable (i.e., displayed an average type-I error rate < 5.0 and power > 20.0) within a nonprimate taxon, in this case the extant heteromyid rodent *Dipodomys*. This study shows that no technique tested displayed a type-I error rate that consistently matched the stated error rate of the analysis while maintaining reasonable power. Therefore, the only techniques that satisfied the empirical criteria were the CV (1% formula and program methods) and Levene's test of relative variation (2.5% method). Of these three methods, the CV program method was slightly more powerful whereas the LEV technique had a lower type-I error rate. On the other hand, the majority of the methods, including the range-based statistics, Levene's 5% technique, and the CV methods that employed 5% significance levels tended to have low average power (< 10) and/or high type-I error rates and should be avoided when testing a single-species hypothesis in heteromyids.

This study also points to a need to select variables carefully. Many workers have suggested the use of upper and lower posterior tooth dimensions (P4/p4–M2/m2), particularly first and second molar dimensions (Gingerich, 1974; Cope and Lacy, 1992, 1995; Cope, 1993). While the results of this study generally confirm these suggestions, there were particular dimensions that had very high type-I error rates (APM2 and tm1) or low power (apm2) across all techniques tested. These differences are likely the result of the different taxonomic groups investigated—previous results were based primarily on primates. Because of these taxonomic differences, a clear understanding of the variation of each variable in the group being studied needs to be obtained prior to employing any of the statistical methods. Dimensions that exhibit a low degree of intraspecific variability and high interspecific variability are probably the most reliable dimensions to use (Cope and Lacy, 1992, 1995). For heteromyid rodents, the widths of the upper second molar (TM2) and lower second molar (tm2) were the most reliable. In addition, while it is not wise to test all dimensions available (to limit increases in the studywise type-I error rate), at least three or four variables (preferably both lengths and widths) should be tested using one of the methods employed here due to the possibility of a type-I error occurring. If only one dimension leads to a rejection of the null hypothesis, conclusions regarding the taxonomic composition of the sample should be tempered. Conversely, several dimensions that lead to a rejection of the null hypothesis provide a solid statistical foundation to conclude that multiple species may be present in a fossil assemblage.

ACKNOWLEDGMENTS

This work was completed as part of my Ph.D. dissertation from Columbia University

under the guidance of Malcolm C. McKenna. I thank Malcolm for always allowing me to pursue my own individual research interests while providing valuable advice and guidance along the way. I also thank L.F. Marcus, J.H. Wahlert, A.D. Barnosky, B.P. Kraatz and R.S. Feranec for their comments and suggestions. My appreciation goes to the following curators and collections managers for allowing me access to specimens under their care: F. Brady and R.D.E. MacPhee (American Museum of Natural History, New York), D.S. Janiger (Los Angeles County Museum, Los Angeles), L. Abraczinskas and P. Hildebrandt (Michigan State University Museum, East Lansing), B. Stein (Museum of Vertebrate Zoology, Berkeley), P. Unitt (San Diego Natural History Museum, San Diego), L.K. Gordon (Smithsonian Institution, Washington, D.C.), and G.D. Baumgardner (Texas Cooperative Wildlife Collection, College Station). I would also like to thank M.G. Lacy for providing copies of his DOS computer programs to test single-species hypotheses. This research was funded in part by a National Science Foundation Graduate Fellowship, a Faculty Fellowship from Columbia University, an American Museum of Natural History Theodore Roosevelt Grant, and a Ford Foundation Dissertation Fellowship.

REFERENCES

Albrecht, G.H., and J.M.A. Miller. 1993. Geographic variation in primates: a review with implications for interpreting fossils. *In* W.H. Kimbel and L.B. Martin (editors), Species, species concepts, and primate evolution: 123–161. New York: Plenum Press.

Barnosky, A.D. 1986. New species of the Miocene rodent *Cupidinimus* (Heteromyidae) and some evolutionary relationships within the genus. Journal of Vertebrate Paleontology 6(1): 46–64.

Cameron, D.W. 1998. Anatomical variability and systematic status of the hominoids currently allocated to the African Dryopithecinae. Homo 49(2): 101–137.

Carrasco, M.A. 1998. Variation and its implications in a population of *Cupidinimus* (Heteromyidae) from Hepburn's Mesa, Montana. Journal of Vertebrate Paleontology 18(2): 391–402.

Carrasco, M.A. 1999. Morphological variation in the dentition of kangaroo rats (genus *Dipodomys*) and its use in distinguishing species in the fossil record. Ph.D. dissertation, Columbia University, New York.

Carrasco, M.A. 2000a. Species discrimination and morphological relationships of kangaroo rats (*Dipodomys*) based on their dentition. Journal of Mammalogy 81(1): 107–122.

Carrasco, M.A. 2000b. Variation in the dentition of kangaroo rats (genus *Dipodomys*) and its implications for the fossil record. Southwestern Naturalist 45(4): 490–507.

Cope, D.A. 1993. Measures of dental variation as indicators of multiple taxa in samples of sympatric *Cercopithecus* species. *In* W.H. Kimbel and L.B. Martin (editors), Species, species concepts, and primate evolution: 211–237. New York: Plenum Press.

Cope, D.A., and M.G. Lacy. 1992. Falsification of a single species hypothesis using the coefficient of variation: a simulation approach. American Journal of Physical Anthropology 89(3): 359–378.

Cope, D.A., and M.G. Lacy. 1995. Comparative application of the coefficient of variation and range-based statistics for assessing the taxonomic composition of fossil samples. Journal of Human Evolution 29(6): 549–576.

Donnelly, S.M., and A. Kramer. 1999. Testing for multiple species in fossil samples: an evaluation and comparison of tests for equal relative variation. American Journal of Physical Anthropology 108(4): 507–529.

Fuller, K. 1996. Analysis of the probability of multiple taxa in a combined sample of Swartkrans and Kromdraai dental material. American Journal of Physical Anthropology 101(3): 429–439.

Gingerich, P.D. 1974. Size variability of the teeth in living mammals and the diagnosis of closely related sympatric fossil species. Journal of Paleontology 48(5): 895–903.

Kelley, J. 1993. Taxonomic implications of sexual dimorphism in *Lufengpithecus*. *In* W.H. Kimbel and L.B. Martin (editors), Species, species concepts, and primate evolution: 429–458. New York: Plenum Press.

Kieser, J.A. 1994. Falsification of a single-species hypothesis by using the coefficient of variation: a critique. American Journal of Physical Anthropology 95(1): 95–97.

Lande, R. 1977. On comparing coefficients of variation. Systematic Zoology 26(2): 214–217.

Martin, L.B., and P. Andrews. 1993. Species recognition in middle Miocene hominoids. *In* W.H. Kimbel and L.B. Martin (editors), Species, species concepts, and primate evolution: 393–427. New York: Plenum Press.

Plavcan, J.M. 1993. Catarrhine dental variability and species recognition in the fossil record. *In*

W.H. Kimbel and L.B. Martin (editors), Species, species concepts, and primate evolution: 239–263. New York: Plenum Press.
Schultz, B.B. 1985. Levene's test for relative variation. Systematic Zoology 34(4): 449–456.
Simpson, G.G., and A. Roe. 1939. Quantitative zoology. New York: McGraw-Hill Book Company, Inc.
Sokal, R.R., and C.A. Braumann. 1980. Significance tests for coefficients of variation and variability profiles. Systematic Zoology 29(1): 50–66.
Wahlert, J.H. 1993. The fossil record. *In* H.H. Genoways and J.H. Brown (editors), Biology of the Heteromyidae. Special Publication American Society of Mammalogists 10: 1–37.
Yablokov, A.V. 1974. Variability of mammals. New Delhi: Amerind Publishing Co. Pvt. Ltd.

APPENDIX 9.1

Specimens Examined

The following is the locality information for the multispecies sample of *Dipodomys* (274 total specimens) from northern Baja California. All specimens were complete skulls. Information is arranged alphabetically by species followed by museum. Abbreviations: AMNH, American Museum of Natural History, New York; LACM, Los Angeles County Museum; MVZ, Museum of Vertebrate Zoology, Berkeley; SD, San Diego Natural History Museum; USNM, United States Natural History Museum; n, sample size.

Dipodomys gravipes (n = 62): Bahía de San Quintín (LACM 33433); Colonia Guerrero, across WNW from Red Cliffs, Santo Domingo entrance (LACM 32111, 32113); near entrance to Santo Domingo Canyon, Colonia Guerrero (LACM 38174); Agua Chiquita, 4 mi. E of San Quintín (MVZ 35704); San Quintín (MVZ 36234; SD 15953; USNM 138910); San Ramón, mouth of Santo Domingo River (MVZ 35655, 35657–35659, 35661–35664, 35666, 36214, 36233); Socorro (MVZ 49860, 49863); Socorro, 20 mi. S of San Quintín (MVZ 49854, 49856, 49858); Santa María near San Quintín (SD 8532, 18513); 3 mi. S of San Telmo (SD 15821, 15822); San Quintín Plain (SD 4997, 4999, 5023–5026, 5035–5037, 22347–22350); Aqua Chiquita Canyon (SD 22346); 1 mi. S of San Ramón (SD 4906); Santo Domingo (SD 4682, 4683, 4704, 4715, 4823, 4885, 4941, 4945, 4946, 4950, 4976, 22351–22356; USNM 245884); mouth of Agua Chiquita Canyon, San Quintín Plain (USNM 245885).

Dipodomys merriami (n = 85): Between El Socorro and El Counsuelo, Hwy. 1, Arroyo San Quintín (LACM 38172); Aqua Chiquita, 4 mi. E of San Quintín (MVZ 35702); Arroyo Nueva York, 15 mi. S of Santo Domingo (MVZ 36244, 36245, 36247); San Quintín (MVZ 49880, 49882–49884, 49886, 49887, 49889–49894; SD 1218, 4995, 4996, 15955, 22207; USNM 138911, 138914, 138915, 138917, 138921, 139827, 139829); Santo Domingo (MVZ 36236, 36237; SD 4668, 4669, 4678–4680, 4684, 4693, 4717, 22210, 22211); 1 mi. S of San Ramón (SD 4912, 4913, 4924, 22209); N end of San Quintín Plain (SD 4937–4939, 4960–4962, 22206, 22208); San Quintín Plain (SD 5042); Santa María near San Quintín (SD 8533, 18519–18521); 7 mi. SE of San Quintín (SD 15800); 10 mi. E of San Quintín (SD 18576–18578, 18587–18589, 18604, 18617, 18619–18622); NE side of San Quintín Bay (SD 19534–19543); 2 mi. S of Old Mill on the N side of San Quintín Bay (SD 19606); near Rock Bluff, 8 mi. N of Cape San Quintín (SD 19607, 19608); 7 mi. N and 0.5 mi. W of Cape San Quintín (SD 20074).

Dipodomys simulans (n = 127): Colonia Guerrero, across WNW from Red Cliffs, Santo Domingo entrance (LACM 32109, 32110); Santo Domingo (LACM 1171; MVZ 36227, 36229; SD 4671, 4673–4675, 4689–4692, 4695, 4698, 4705–4708, 4714, 4886, 4901, 4903, 22335–22341); Valladares (MVZ 35681–35687); San Antonio Ranch, Santo Domingo River (MVZ 35689); San José (MVZ 35695, 35696, 36073, 36075–36082); Colnett (MVZ 36088); San Quintín (MVZ 49831, 49832; SD 4992, 5003, 5005, 5007, 5030, 5031, 15956, 15957, 22333, 22334; USNM 138909, 139819, 139820, 139822, 139823, 139826, 245886); San Telmo (MVZ 35668–35678, 35680, 35921, 36219, 36220, 36222, 36223–36225; USNM 139830); Socorro, 20 mi. S of San Quintín (MVZ 49861, 49862); 2.4 km S and 5.0 km W of Mission Santo Domingo (MVZ 153965, 153966, 153968, 153969); 2.4 km W of Mission Santo Domingo (MVZ 153960–153964); 5 mi. W and 1.25 mi. S of San Telmo de Abajo (MVZ 148091–148100); 1 mi. S of San Ramón (SD 4907, 4925); N end of San Quintín Plain (SD 4958); Santa María near San Quintín (SD 18514); 10 mi. E of San Quintín (SD 18585, 18603, 18615, 18616); NE side of San Quintín Bay (SD 19533); near Rock Bluff, 8 mi. N of Cape San Quintín (SD 19605); Aqua Chiquita Canyon near San Quintín (SD 22327–22332); 20 mi. SE of San Telmo (USNM 528823).