AMERICAN MUSEUM
Novitates

PUBLISHED BY
THE AMERICAN MUSEUM
OF NATURAL HISTORY
CENTRAL PARK WEST AT 79TH STREET
NEW YORK, N.Y. 10024 U.S.A.

NUMBER 2584 AUGUST 19, 1975

SYDNEY ANDERSON

On the Number of Categories in Biological Classifications

AMERICAN MUSEUM
Novitates

PUBLISHED BY THE AMERICAN MUSEUM OF NATURAL HISTORY
CENTRAL PARK WEST AT 79TH STREET, NEW YORK, N.Y. 10024
Number 2584, pp. 1-9, figs. 1-6. August 19, 1975

On the Number of Categories in Biological Classifications

SYDNEY ANDERSON[1]

ABSTRACT

The theoretical maximum and minimum numbers, and the most probable numbers, of categories to be recognized in classifications designed to express all cladistic information in groups of different sizes are derived by Monte Carlo models based on a theoretical distribution that fits real taxonomic data. The number of categories required is much nearer the minimum possible number than the maximum possible; usually 11 to 16 categories will be needed for a group of 100, 21 to 26 for a group of 1000, and 26 to 36 for a group of 10,000. The number of categories required for a group of a certain size increases as the percentage of the members of that group that are extinct increases.

INTRODUCTION

The number of categories employed in hierarchies of biological classifications has been basically arbitrary. It has been historical or traditional, rather than justified in any explicit or theoretical sense. Linnaeus used empire, kingdom, class, order, genus, species, and variety. The principal subsequent changes were adding family and phylum.

An author may have a feeling for an optimal size of a taxon. If a certain size is desired, the number of categories must be increased as the total number of recognized species increases. Size here means the included number of taxa of the next lower category. The concept of an optimal size is also arbitrary, arising probably more from considerations of mnemonic convenience than from theoretical premises. In practice, taxa are not of equal sizes, whether phenetic, eclectic, or cladistic criteria are employed in establishing a classification (Anderson, 1974b).

If no limit is assumed on the number of species that a single species could split into simultaneously, the range of possible values (x) for the maximum number of successive splits in any one line within any tree with n terminal twigs (species) would be some value from 1 through n−1; 1 if a single "explosion" occurred, and n−1 if all splits were in a single line.

If only binary splits are assumed, the value for x would be an integer not less than log n/log 2, based on the relationship $2^x = n$, and not more than n−1. For 10^6 species, x would be at least

[1]Curator, Department of Mammalogy, the American Museum of Natural History.

ISSN 0003-0082 / Price 70 cents

log (10^6)/log 2, or 6/0.30103, which to the next integer equals 20, and would not be more than 999,999. These are boundary conditions for x, which value also is one less than the number of categories (i.e., ranks or levels in a hierarchy) required to express all branches of a phylogenetic tree in the classification thereof. The pattern that prevails in nature, however, lies at neither extreme.

The prevailing pattern is not significantly different from the pattern postulated on the basis of a null hypothesis that the evolutionary events of speciation (or splits of lineages) and extinction (or termination of lineages) occur randomly both as to the twigs hit by the event and the time of the event (Anderson, 1974b; Anderson and Anderson, 1975). Any sample of a fauna or an ecological community that differs from the null hypothesis, and there are comparatively few of these, is of special interest.

MODEL WITHOUT EXTINCTION

In figure 1 are plotted three different values for the greatest number of consecutive binary splits in any one lineage within trees with different numbers of terminal twigs. The three values are the maximum possible number, the minimum possible number, and the average number. The average number is based on the assumption that only splits occurred, that no species (twig) became extinct during the run, and that

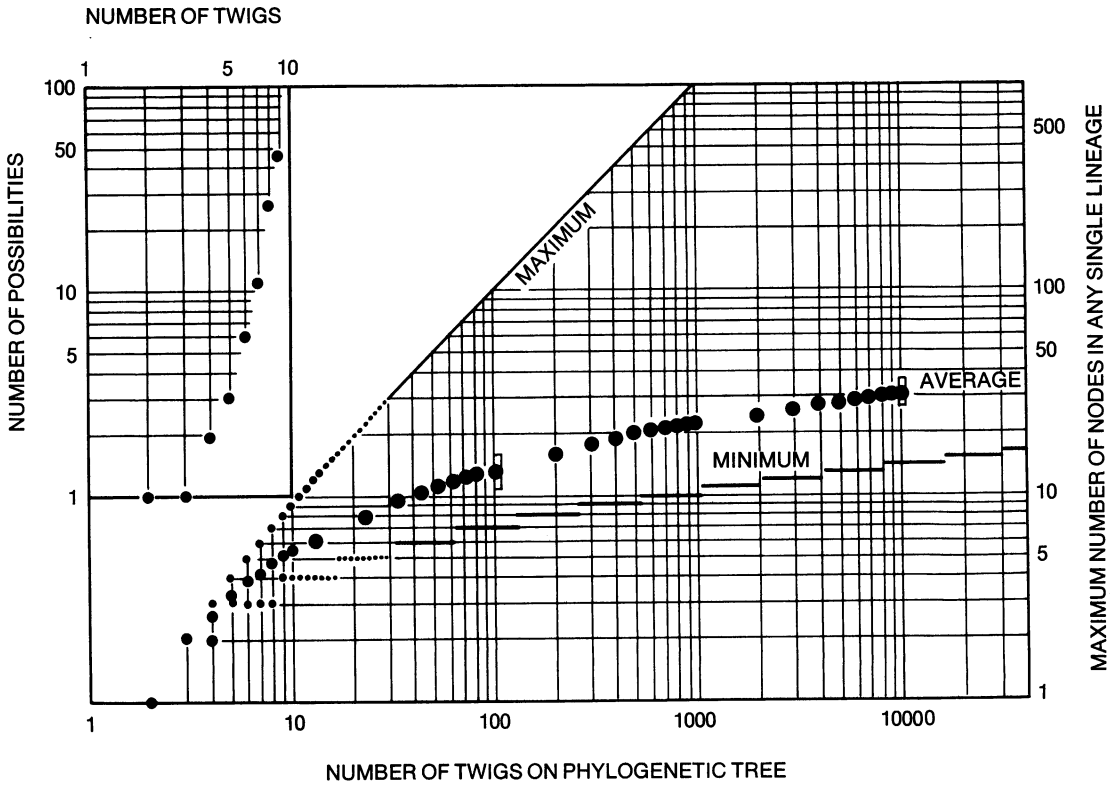

FIG. 1. Graph (upper left) showing rapid exponential increase in numbers of possible trees with different numbers of twigs, from two to nine. Larger graph shows numbers of nodes in lineage with most nodes in the tree, which number is one less than number of categories required to classify trees with various numbers of twigs. Maximum and minimum possible values and average values (dots) for 10 trees generated by a computer as described in text. For trees of 100 and 10,000 twigs, values one standard deviation above and below the mean are marked.

the probability of a split occurring was equal for all species. The values up to a tree of 10 species were calculated by iterating all possibilities and the probabilities of each and summing them for trees of each size. Average values above 10 are based on 10 runs of a Monte Carlo model. I did not think of any way to compute them directly on a basis of probabilities. The possibilities increase geometrically, and the iterative approach employed for trees with up to 10 twigs became too tedious to pursue further.

The Monte Carlo method may briefly be described as the study of an artificial stochastic model of a physical or mathematical process. When an equation or relationship arising in a nonprobabilistic context demands a numerical solution not easily obtainable by standard numerical methods there may exist a stochastic process or model with distributions which satisfy the equation and it may be more efficient to construct such a model and derive the solution from it than to attempt a standard numerical solution.

The number of possible forms of a tree with eight twigs is 26, for a tree of nine twigs the number is 47 (see fig. 1). The precise form of the tree will not be dealt with hereafter. Only frequency distributions of twigs with different numbers of branching points (nodes) in their lineages are discussed. The number of nodes in the lineage with the most nodes is given special consideration.

The first draft of the computer program for the first Monte Carlo model to be discussed (MODEL.04) was written by Charles S. Anderson, who also assisted in the debugging of MODEL.06. Debugging and completion of the program was done by the author. The programs (MODEL.04 and MODEL.06) were in BASIC language and were run on a PDP-8/E computer at the American Museum of Natural History.

The model singled out a twig randomly and split it. This means that the end of the twig was assumed to have grown into two twigs, the base remaining as a discrete segment between two nodes of the tree. A count was kept of the number of prior events and the numbers of twigs having different numbers of branching points or nodes in their lineages, and at selected intervals these values were printed out. The only value of current interest is the number of nodes in the lineage with the most such points, which value, as noted before, is one less than the number of categories required if every rank is recognized as a category in a classification.

Ten runs were made of this Monte Carlo model. The values derived for the number of categories needed for a tree of a certain size are not normally distributed. The distribution is skewed so that more of the values lie near the lower end of the range. The means for trees of several selected sizes in the 10 runs were calculated and plotted in figure 1. The range for plus and minus one standard deviation is plotted for groups of size 100 and 10,000.

It is apparent from figure 1 that the most probable value for the number of categories required is consistently nearer to the minimum limit than to the maximum limit after the tree has nine twigs. The more twigs on the tree the smaller the percentage of the minimum to maximum range that lies above the average.

Projecting the curve for probable numbers of categories required for a group of 10 million, which is about the number of all living animal species according to some estimates, it seems that about 70 categories would be required. The number of species actually described up to now is, of course, much less, perhaps in the neighborhood of 1.5 to two million, and about 60 categories would be required for these.

MODEL WITH EXTINCTION

The above model assumes no extinct species and, therefore, resembles a situation where only the living species of a group are being classified. The effect of extinction on the number of categories was simulated by another model (MODEL.06) as follows: A tree of a specified size (i.e., number of twigs) was generated as above, then a decision was made by the computer on a random basis whether to terminate or split a twig randomly selected. No terminated twig (or extinct species) could be selected for a later split. Probabilities of extinctions and splits were made equal so the number of species (twigs) living would tend to remain constant as successive evolutionary events occurred. In this model there are two kinds of evolutionary events, splits and extinctions. The percentage of twigs that were extinct

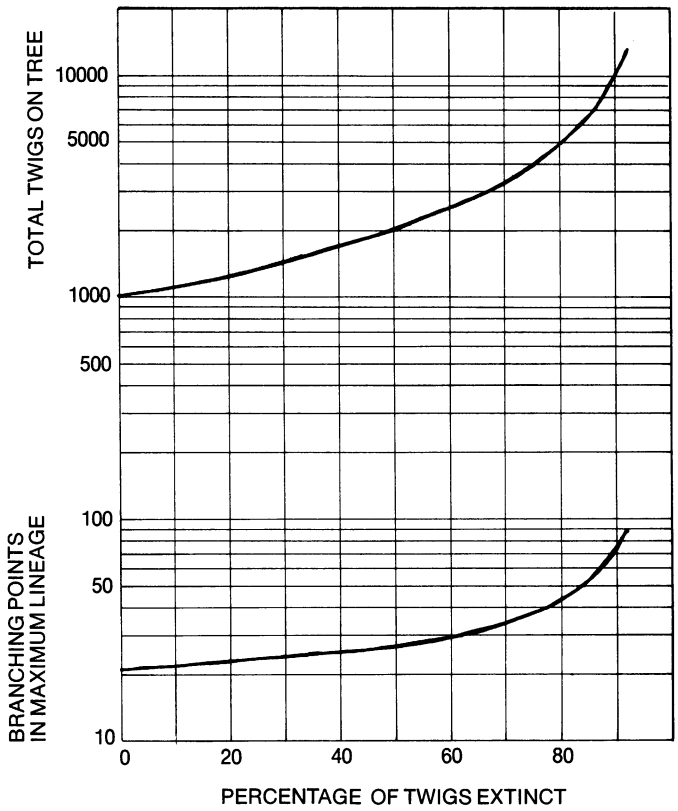

FIG. 2. Graph showing size of a tree with 1000 living twigs as the percentage of extinct twigs is increased up to 92 percent of all twigs on tree (above) and (below) number of branching points in longest lineage (i.e., lineage with most branching points), both based on a single computer run. A second run, up to 80 percent extinction, gave virtually identical results.

would increase with time. In this way we can see how the number of categories needed to express the complete cladogram in a classification would change as the extinct species increased in number relative to the total number of species living and extinct.

In figure 2 the results of a computer run are plotted. The possibility of extinctions was introduced at a tree size of 1000 and the run continued up to 25,000 events, at which time the tree had about 12,000 twigs, 1000 living and 11,000 extinct. The number of categories required is one more than the branching points in the maximum lineage, which is plotted also. The number 1000, for living members of the group, is not only a conveniently visualized round number but is about the number of genera of living therian mammals. The number of extinct genera of therians is somewhere between two and three thousand, so the percentage of extinct genera in the total fauna (or terminal branches on the phylogenetic tree, disregarding the subgeneric parts thereof, for the moment) lies between 66 and 75 percent. The graph suggests, therefore, that about 32 to 37 categories would be needed to classify this phylogeny, under the conditions stipulated above.

The program for MODEL.06 allows the operator to specify when extinctions are to begin and what tree sizes are to be generated. I would be glad to simulate a fauna and predict the number of categories required for a group if the numbers

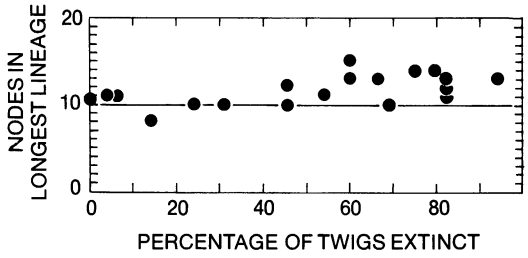

FIG. 3. Graph showing longest lineages in trees with 50 twigs, different percentages of which are extinct. Points are values derived in each of 18 Monte Carlo simulations. Point for no extinctions is the mean of 10 runs from figure 1.

of living and extinct members are specified. Or, I will provide a copy of the program if anyone wants to have it.

Comparison of the lengths of (i.e., number of branching points in) the maximum lineages in figures 1, 2, and 3 shows that more categories will be required in classifying a group of a given size with extinct members than a group of the same size without extinct members. Furthermore, the greater the percentage of extinct members in a group of a given size, the greater the number of categories required. A group of 10 thousand members of which 90 percent are extinct will require as many categories as a group of 10 million in which none are extinct. This is for the obvious reason that splits do not occur in extinct species, but are confined to the ongoing living part of the tree. The other assumptions of the model should be remembered here, such as the assumption that the probability of an event is equal in all lines in which it may occur. The model is stochastic. We are dealing with probabilities rather than certainties. It is possible in a specific run of the model for the lineage with the most branching points to be extinct, but in actual runs this rarely occurs (i.e., fewer than one time out of 20). I predict that in the best models of phylogenies we will be able to develop it will rarely be found to have occurred.

The models described give us some theoretical values for numbers of categories required under the assumption that all clades in a phylogenetic tree are to be expressed in the classification and suggest that the number of categories needed may not be so great as might be supposed, although the number is greater than taxonomists have customarily used.

I do not advocate the inclusion of all topological relationships of the cladistic diagram in the formal set of named taxa. This is a consideration I have briefly discussed before (Anderson, 1974a).

It should be mentioned here that the method of constructing cladistic diagrams may result in showing branches at places where none occurred in the actual phylogeny from which the sample being analyzed was drawn. For example, two samples drawn from a single segment (i.e., a lineage between two nodes) of the tree at points between which differences arose would be treated as sister groups in the analysis.

I suggest that the analysis should go beyond that level, when fossils are concerned at least. Once the synapomorphies are established with sufficient confidence and the cladistic diagram of sister groups is drawn, the diagram may profitably be modified by introducing a time scale so far as information will allow. Then those cases in which a taxon (1) existed early enough in time to have been an ancestor of another taxon, (2) exhibits the traits expected in an ancestor, based on the earlier analysis, and (3) occurs where the ancestor probably occurred, should, in the interests of parsimony, be so regarded. This, of course, alters the form of the model upon which one may wish to base a classification.

Once these modifications have been made in the cladistic diagram, we have a model of the phylogeny. It will be a good or bad model, that is the closeness of its resemblance to the real phylogeny will vary, depending chiefly on our knowledge.

After the best possible phylogenetic model has been constructed, the question of how best to classify should be addressed.

Let us consider a hypothetical example.

Suppose we have a presumably monophyletic group of 11 taxa, for which a cladistic model is diagramed as shown in figure 4. The maximum of nodes in any one lineage is seven, so eight categories, including the initial category, are required for a complete classification (i.e., one in which each clade is a named taxon and each rank is a category). The total number of implied clades is 21. A clade is defined as any segment of the

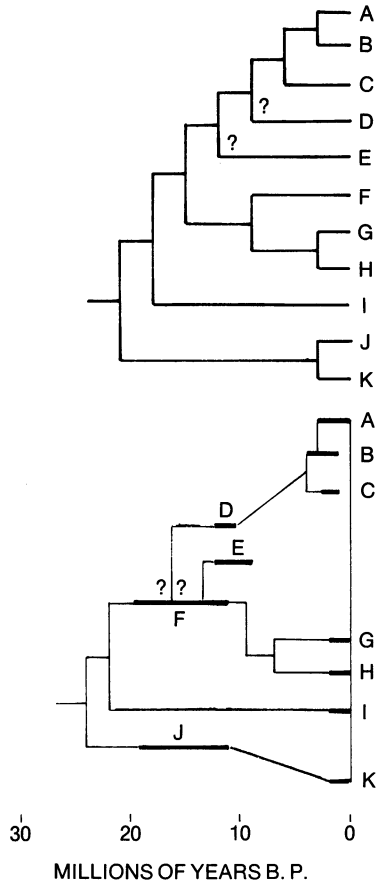

FIG. 4. Hypothetical group showing relationships by a cladistic diagram above and a phylogenetic diagram below. The phylogenetic diagram is derived from the cladogram by incorporating time and possible ancestral relationships as discussed in text.

diagram and all of its subsequent branches. Since every segment begins a clade, the number of clades is equal to the number of segments. Now suppose that the time scales for known specimens of each of these taxa are added and it is noted that B could have been ancestral to A; D ancestral to A, B, and C; F ancestral to G and H, and possibly to A, B, C, D, and E; and J ancestral to K. The resulting phylogenetic model is also shown in figure 4. The maximum of nodes is now five; so six categories in all are needed. The number of clades is now 14.

The above comparison suggests that the number of categories that might be used depends upon the conventions of the method of analysis, among other things. (The example, incidentally, is a simplified presentation of the present understanding of relationships of genera within the superfamily Hominoidea, including the apes and gibbons.)

SOME EXAMPLES FROM MAMMALIAN CLASSIFICATION

Let us now consider two classifications within the order Primates. In figure 5 is shown the classification of Primates as it was arranged in a card file by McKenna et al., here at the American Museum of Natural History several years ago. The twigs are genera or groups of genera (the number of genera in a group is indicated in parentheses after the name of one of the genera included). I have worked back from the classification to construct the diagram of relationships. I thus infer monophyletic groups where those who arranged the classification may not have intended to imply this in all cases. There are still a number of multichotomies that should be examined and resolved into dichotomies as far as possible. In drawing this presumed (or *ex post facto*) cladogram, I have adopted the convention of placing a branch with fewer twigs above one with more. The twigs are the groups of genera or single genera shown. The category of each named taxon is shown by one or more letters; ST = subtribe, T = tribe, SF = subfamily, F = family, SRF = superfamily, IO = infraorder, SO = suborder. The lineages with distal living members are shown with darker lines. The general arrangement of these living lineages does not differ much from that outlined in the 1910 edition of the Encyclopedia Britannica by Richard Lydekker. The major differences are (1) the placement of *Tarsius* with the haplorhines rather than the lemurs, (2) the inclusion explicitly of a larger number of extinct genera, mostly described since 1910, (3) the separation of *Callimico* from the other marmosets, (4) the movement of a few taxa from one category to another of higher or lower rank, and (5) using fewer implied ranks within the cebids.

The greatest number of ranks from generic group through order is seven, although eight

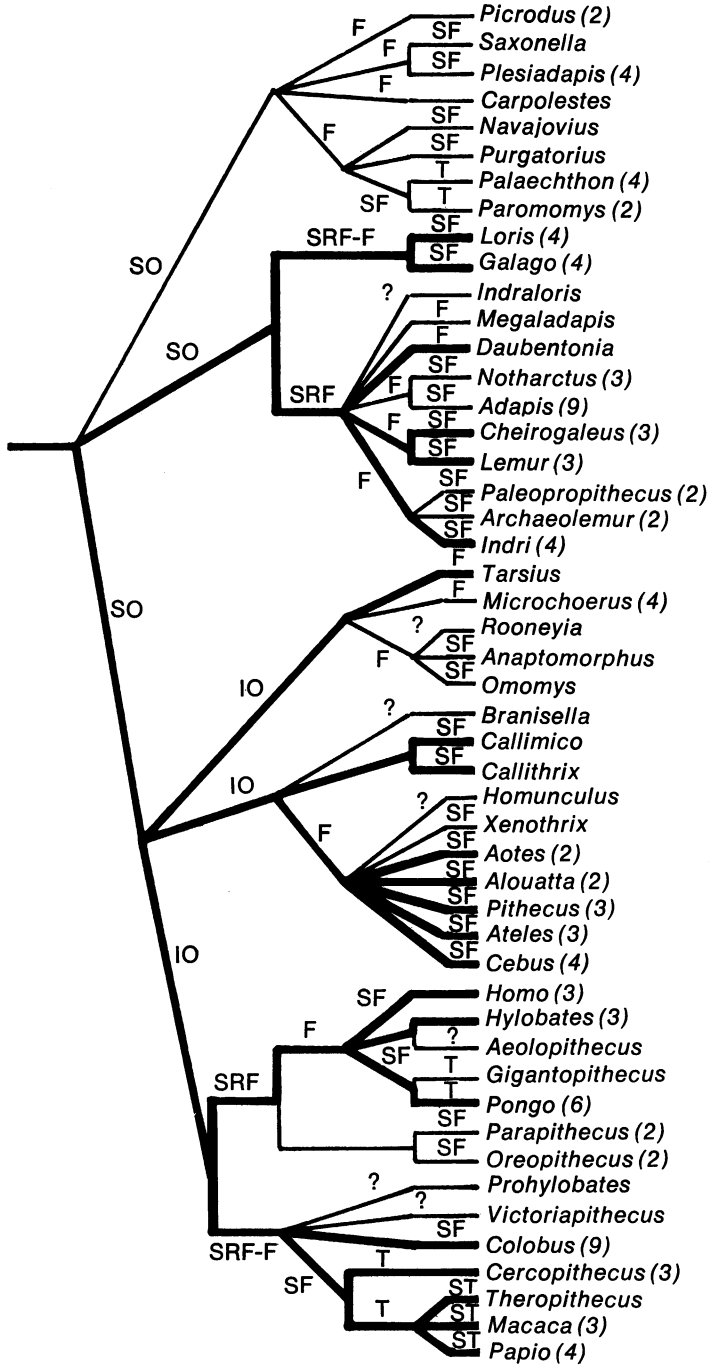

FIG. 5. Cladogram of primate relationships down to level of generic group (from card file of McKenna et al.).

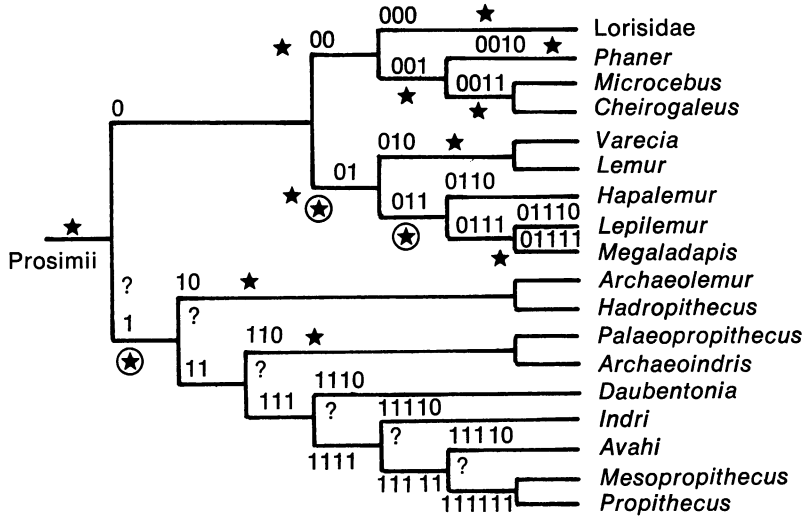

FIG. 6. Cladogram of the relationships of Malagasy lemurs from Tattersall and Schwartz (1974, redrawn). Stars indicate those clades that are given formal names in their classification. Stars in circles indicate paraphyletic taxa. Several relatively less certain connections are indicated by question marks.

named categories are used in the classification. There are 49 generic groups, of which 27 are extinct. My model predicts 12 categories for a group of 50 with 55 percent extinct, when only dichotomies exist. In this classification there are seven rather than 12 categories because of the unresolved multichotomies. The model indicates 12 as the most probable value, not as the only possible value. In any event, the agreement of model and data is quite close.

A recent application of cladistic analysis to one major group of primates is that of Tattersall and Schwartz (1974). Their cladogram (redrawn) for the Malagasy lemurs is shown in figure 6 (present paper). The convention of placing a branch with fewer twigs above one with more twigs was used by me in arranging the diagram, and clades have been numbered for convenience of discussion and as a further illustration of one method of expressing relationships. Binary numbers are used for dichotomies, the upper branch is 0 and the lower is 1.

The provisional classification of Tattersall and Schwartz does not attempt to employ exclusively monophyletic (i.e., holophyletic) taxa. For example, clade 1 includes a clade, 1110, recognized as a family. The family Indriidae (clade 1 less clade 1110) is, therefore, paraphyletic. Likewise, clade 01 includes clade 011 11, recognized as the family Megaladapidae. The family Lemuridae (01 less 011 11) is a paraphyletic taxon, as is the subfamily Lepilemurinae (clade 011 less 011 11). A number of clades have no names in the classification, these clades are 0, 0111, 11, 111, 111 11, and 111 111. I don't recommend giving them names. The authors coined informal names in some cases. Clade 0 is the "lemur/loris group." Clade 0011 is referred to somewhat awkwardly as the "common ancestor of *Microcebus* and *Cheirogaleus* [and] . . . both modern genera" in text (p. 184), and it is called "tribe Cheirogaleini" in the classification. Incidentally, the name Cheirogaleinae is used in text where, presumably, Cheirogaleidae is meant. The name Cheirogaleinae is not in the classification on page 188. The basal segment of clade 0111 is called the "common ancestor of *Lepilemur* and *Megaladapis*." Clade 1 is called the "*Indri*-group" in text and the "infraorder Indriiformes" in the classification. In the cladogram, several infraorder connections that were regarded by Tattersall and Schwartz as less certain than most were shown

by them with broken lines and I have shown these (in fig. 6) with question marks.

The 18 genera (including the Lorisidae, which were not subdivided in the cladogram, as one) exhibit eight ranks or categories, which is the most probable or predicted number for a group of 18, with six extinct twigs, according to my model.

CONCLUSION

Frequency distributions of taxonomic groups of organisms usually are not significantly different from the distribution resulting from a null hypothesis that the evolutionary events of extinction and splitting of species occur randomly. A Monte Carlo model based on this assumption enables us to predict the maximum numbers of ranks that will probably occur in a cladistic diagram in which all splits are dichotomous for a fauna or flora of any size and for different percentages of extinction.

Two examples from mammalian classification are presented and in both cases the prediction of number of ranks was correct. As further cases are worked out in detail, it will be interesting to see how well the model fits them.

The conventions employed in constructing a classification from a cladogram will affect the number of ranks therein that are designated as formally named categories, as well as which ranks therein are designated as which categories. Different taxonomists use different conventions, as does one taxonomist at different times, and, therefore, the conventions or rules being used in each classification should be explicitly stated. This assumes that taxonomists should communicate with each other and with other people.

LITERATURE CITED

Anderson, Sydney
1974a. Some suggested concepts for improving taxonomic dialogue. Syst. Zool., vol. 23, pp. 58-70, figs. 1, 2.
1974b. Patterns of faunal evolution. Quart. Rev. Biol., vol. 49, pp. 311-332, figs. 1-15.

Anderson, Sydney, and Charles S. Anderson
1975. Three Monte Carlo models of faunal evolution. Amer. Mus. Novitates, no. 2563, pp. 1-6, figs. 1-3.

Tattersall, Ian, and Jeffrey H. Schwartz
1974. Craniodental morphology and the systematics of the Malagasy lemurs (Primates, Prosimii). Anthrop. Papers Amer. Mus. Nat. Hist., vol. 52, pp. 139-192, figs. 1-24, tables 1-3.