Meehl, P. E. (1983). Consistency tests in estimating the completeness of the fossil record: A neo-Popperian approach to statistical paleontology." In J. Earman (Ed.), *Minnesota studies in the philosophy of science: Vol. X, Testing scientific theories* (pp. 413-473). Minneapolis: University of Minnesota Press.

Consistency Tests in Estimating the Completeness of the Fossil Record: A Neo-Popperian Approach to Statistical Paleontology

Paul E. Meehl

Most educated persons today, who have been repeatedly exposed since childhood to pictures of the famous "horse series," are likely to think of it, or of the fossil dinosaurs they have seen in museums, rather than other lines of evidence for the theory of evolution. They differ in this respect from Charles Darwin, who did not consider the fossil data to be supportive of his theory. On the contrary, he viewed the paleontological findings available in his day as perhaps presenting "the gravest objection which can be urged against my theory" (Darwin 1859, p. 280), emphasizing the fossil record's failure to provide the "infinitely numerous transitional links" (p. 310) that would illustrate the "slow and steady operation of natural selection" (Eldredge and Gould 1972, p. 87). Because he thought of the fossil record as adverse rather than supportive, his chapter on the record is appropriately titled by reference to its *imperfection,* and appears toward the end of the book, following the general chapter on "difficulties of the theory."

Philosophers of science, and psychologists of certain persuasions (eg., Freudians who wish to defend the "scientific" status of psychoanalytic doctrine against super-operational critics), have sometimes invoked the theory of organic evolution as a paradigm case of a theory that is generally admitted to be empirical and scientific but that mostly offers after-the-fact explanations of data without being able to make predictions of those data. One thinks of evolutionary theory as explaining that there is such a creature as the rhinoceros, but nobody has ever claimed, given the general postulates of evolutionary theory, to *derive* the empirical statement that an animal of such-and-such rhinoceroid properties would evolve. Sir Karl Popper (1974, pp. 133-143), with his emphasis on falsifiability as a criterion of scientific theories, has gone so far as to suggest that we should look upon Darwinism (he specifically includes neo-Darwinism) not as a scientific theory at all, in the strong sense he uses the term "theory," but as a metaphysical speculation that has been fruitful in science.[1]

Few of us find that characterization satisfactory. Aside from the philosophical puzzle of just how a metaphysical theory can be "fruitful" if it (literally) has no factual consequences, however vague or "weakly implied, most biological scientists could not rest content with Poppers diagnosis of the theory of organic evolution as not a scientific hypothesis but merely a fruitful metaphysical speculation. We must parse the issues, separating the well-known disputes about whether the notion of adaptation can be defined

[1] Professor Malcolm Kottler has kindly called my attention to Popper's article "Natural selection and the emergence of mind" (*Dialectica*, *32*, 1978, pp. 339-355), in which Sir Karl modifies his long-held view to this extent, that it is possible to reformulate Darwinism in such a strong fashion that it does become testable; in which strong form it is, he says, known to be false. Evolutionists will hardly be happier with this amended view than with his earlier one, and it will be clear to readers that the amendment does not nullify our original problem or the point of this paper.

adequately without a vicious circularity (not all circularities are vicious!), and whether it is possible to explicate a logically adequate notion of after-the-fact explanation in a theoretical structure that would not have permitted before-the-fact prediction; and setting aside what factors other than slight selection pressures may play a critical role (e.g., founder effects, genetic drift, geographical isolation). Then we might hold that there is a minimal content of evolutionary theory that is falsifiable in principle, and hence would escape Popper's classification of the theory as metaphysical, to wit: There must have existed a large number of intermediate transitional forms, even under the modified hypotheses of neo-Darwinism. The probandum should, at least in principle, be falsifiable by the fossil record, *if suitable bounds can be put on find-probabilities.*

Even during the ascendancy of the somewhat dogmatic and overly restrictive Logical Positivism (Vienna Circle of the 1920s), everyone recognized the necessity to distinguish between the technical feasibility of confirming a scientific theory and the abstract possibility of doing so "in principle." Early on, there were controversies about how to construe "in principle." For example, attackers of the verifiability criterion of meaning made the telling point that unless the phrase "in principle" were construed so narrowly as to be nearly the same as "technically" (with present knowledge and instruments), one had to allow for the future possibility of novel auxiliary hypotheses, perhaps taken together with to-be-developed scientific instruments or analytic procedures. But allowing for such future possibilities meant that to deny the confirmability of the theory "in principle" required one to deny, in advance and a priori, the possibility of somebody's being clever enough to cook up the appropriate theoretical auxiliaries and associated technological instruments. This advance general proof of a negative would seem to be, on the face of it, an impossible intellectual feat.

The relevance of this old controversy to the present problem is obvious: One way to escape the charge of nonfalsifiability against evolutionary theory is to sketch out, in as much detail as the present state of knowledge permits, how one would go about estimating the numerical adequacy of the fossil record. This estimation is a necessary stage in evaluating the numerous intermediate probanda of the theory (eg., the hypothesis that between the extant species of whales and their presumed land mammal ancestor, there existed numerous protowhale transitional forms). It is not necessary, in order to escape the charge of untestability, for us to have *presently* available all of the kinds of data that would generate good estimates of the quantities that we conjecture will converge numerically. The philosophical point is that there is nothing unfeasible in principle about gathering such data, and in fact some of them, such as the dates at which the first finds of a given extinct species were made, are part of standard paleontological practice. I am prepared to argue that to the extent the completeness of the record remains an interesting live question (or, in some rare quarters, a persisting basis for doubt as to the macro-evolutionary hypothesis itself), then if the mathematics I suggest are essentially sound, this provides motivation for adopting whatever practices of recording, cataloging, and pooling information are necessary in order for the first step of raw data classification required by my proposed methods to occur. But even if that motivation were inadequate, I should think I had provided an answer to the Popperian complaint of empirical nonfalsifiability.

With the passage of time and the subsidence of theological and philosophical controversies, contemporary biologists and geologists can afford to take a detached and relaxed view, now that the overall concept of major evolution has become part of the mental

furniture of almost all educated persons. The imperfection of the geological record is taken for granted, and we have causal and statistical considerations explaining why the record should be expected to be incomplete. Nevertheless the dearth of really good fossil sequences has remained a mild nagging source of intellectual discomfort, and from time to time some convinced evolutionist finds himself in hot water with his colleagues in dealing with it. Thus Goldschmidt (1940) was so impressed by this absence of transitional forms that when he combined it with skepticism about the accumulated micro-mutation *mechanism* of neo-Darwinism, he was led to propound an alternative idea of extensive chromosomal changes. This idea was criticized by Dobzhansky (1940, p. 358) as involving as much of a "miracle" as out-and-out mystical extra-scientific hypotheses that Dobzhansky knew Goldschmidt would not have countenanced.

Quotations abound, and since universal knowledge of the epistemic role assigned to the record's imperfection may be presumed, I shall not bore the reader by piling up such quotations. I confine myself to mentioning a recent alternative theory, "Punctuated Equilibria: An Alternative to Phyletic Gradualism," set forth by Miles Eldredge and Stephen Jay Gould, largely motivated by the missing transitional forms problem. They suggest that many breaks in the fossil record are "real" (1972, p. 84), that the interpretation of the incompleteness of the record is unduly colored by the paleontologists' imposition of the theory of phyletic gradualism upon otherwise recalcitrant facts, and that the rarity of transitional series "remains as our persistent *bugbear*" (p. 90). They refer to the reputable claims of Cuvier or Agassiz, as well as the gibes of modern cranks and fundamentalists, for whom the rarity of transitional series has stood as the bulwark of anti-evolutionists arguments (p. 90). They also say that "we suspect that this record is much better (or at least much richer in optimal cases) than tradition dictates" (p. 97).

The theory-demanded assertion of extreme record incompleteness, despite the fact that it no longer constitutes a serious objection to the essential truth of the theory of evolution in the minds of most scientists, does, as these contemporary authors indicate, still persist as a theoretical irritant, so that one would prefer, if possible, to have something more detailed and definite to say. It could also be argued, without tendentiousness, that for a science that is essentially "historical," having as its aim to reconstruct the unobserved past through a study of its material residues available to us today, the *quantitative degree of incompleteness* of the fossil record must surely be rated a "Big Parameter." Such a Big Parameter is so important theoretically that even very rough, crude estimates, putting bounds on orders of magnitude, are worth having if such could be achieved.

How poor is the fossil record? Unfortunately it is easy to present plausible arguments on both sides, and such counterposed armchair arguments (I do not use the word "armchair" invidiously, since, as Bertrand Russell once said against the American behaviorists, an armchair is a good place to think!) can lead us to expect that the record would be fairly complete, at least complete enough to help answer some of the big questions, such as whether evolution took place gradually or by saltation; or the arguments can persuade us that the record is surely far too scant, *and must remain so*, for any such inferences. During the time that the substantial correctness of the theory of evolution itself was still a matter of dispute among educated persons, it was easy to find these kinds of considerations and counterconsiderations in the polemical literature. Nor have these old arguments on both sides been shown to be qualitatively invalid in the meantime. It is merely

that the acceptance of the evolutionary paradigm has greatly reduced their importance, and therefore we do not find it easy to work up much interest in them any longer.

To summarize briefly the countervailing arguments: On one side, the evolutionists pointed out the unlikelihood of an animal specimen, even one with hard parts, dying in a situation in which the ravages of wind and rain and scavenger animals would leave it intact; then the unlikelihood, even if the animal escaped such quick destruction, that it would form a fossil; then the improbability that it would remain undisturbed by geological faulting, extreme temperatures, water erosion, and so forth; and finally the improbability that even an army of industrious geologists and paleontologists would happen to dig at just the right place. Although these arguments are not in quantitative form and were originally offered as ad hoc explanations of the scanty phylogenies (compared with what the theory might at first have led us to expect), we can rely on a quote from a recent historian of science that warns, "do not make a mockery of honest ad hockery.

Anti-evolutionists, admitting these considerations, were in the habit of rebutting them by saying that, although we don't expect any individual saber-tooth tiger to have been preserved from destruction and fossilized and dug up—a sequence admittedly of very low probability—nevertheless the saber-tooth tiger species was extant for many millions of years; at any one time, hundreds of thousands of tigers roamed the earth; so the net expected number of fossil finds is still "good enough," because we are multiplying a very big number by a very small number.

Of course the trouble with these arguments, rebuttals, and rejoinders is that although they are all intrinsically plausible as *qualitative* (or weakly "semi-quantitative") arguments, they do not lead to even a rough *numerical* value for anything. Consequently the role of the dominant theoretical paradigm is even more crucial in determining whether one finds them persuasive in his subjective, personalist probabilities than in other branches of scientific controversy (cf. Eldredge and Gould, pp. 83-86). In this essay I offer some tentative and rather general quantifying suggestions, which, despite their generality and the idealizations made, do advance us beyond the purely qualitative con–siderations found in the controversy's history. Formulas are derived that, if the methods are sufficiently robust under real departures from the idealization, provide *numerical estimates that we may hope will converge*.

Because of the novelty of my approach and the weight of received doctrine that speaks prima facie against it, I think it necessary to introduce certain general methodological considerations (stemming from a neo-Popperian philosophy of science) as a framework within which the more specific statistical suggestions are set forth. I do not attempt to "establish" or "justify" these methodological guidelines, but merely set them out so that the reader will understand the metatheoretical framework in which my statistical proposals find their place.

1. When qualitatively persuasive arguments of the kind quoted above are in such collision, and when rebuttals to them on the respective sides are not dispositive, so that subjective or personalistic probabilities (especially so-called subjective priors) are the basis of a scientist's choice, it is desirable, whenever possible, to reformulate such arguments in a quantitative rather than a qualitative form. If the state of theory and factual knowledge is not strong enough to permit that, it may be possible to express the expected *consequences* of these qualitative arguments in a quantitative form. That is the step attempted in this paper.

2. Quantitative treatment in all sciences, but especially the life sciences, unavoidably relies upon substantive idealizations in the embedding theoretical text and, either springing from these or added to them, mathematical idealizations in the formalism.

3. In the early stages of efforts to quantify a hitherto qualitative enterprise, roughness in numerical values arising from measurement and sampling errors, as well as from the idealizations in the text and formalism mentioned above, will usually be unavoidable. One hopes that these crude values can be refined later; but even as a first step, before refinement, such numerical estimates are scientifically preferable to a purely impressionistic subjective evaluation of nonquantitative arguments, however qualitatively meritorious.

4. Reasonably approximate numerical point values and reasonable tolerances of numerical error are almost always preferable scientifically to so-called exact significance tests or even exact confidence intervals, neither of these even being really exact anyhow, for reasons well known to statisticians (Meehl 1978, Morrison and Henkel 1972).

5. Numerical agreement, within reasonable tolerances, of two or more nonredundant modes of estimating a theoretical quantity is almost always scientifically preferable to single values estimated by only one epistemic route, even if the latter are accompanied by so-called exact standard errors.

6. The evidentiary weight with regard to a theoretically important quantity provided by several nonredundant convergent estimates increases with their diversity far more than replicated estimates relying upon the same evidentiary path (instrument or procedure of estimation).

7. Although we commonly speak of "hypotheses" in contrast to "required (auxiliary) assumptions," there is no logical or epistemological difference between the substantive theory of interest and the auxiliary assumptions, when both play the same role in a conjunction of assertions to explain or predict observational data. The only difference between what are usually called "hypotheses" (referring to the substantive theory of interest) and "assumptions" (referring to those auxiliary conjectures employed in the formalism or embedding text when testing the substantive theory) lies in the focus of the investigator's current interest. He may be treating only one of them as problematic, although if pressed, he would usually be willing to concede that both are somewhat so. In their logical, mathematical, and evidentiary status, the hypotheses and the auxiliary assumptions are often interchangeable.

8. A difference, however, sometimes exists when the auxiliary assumption has itself already been well corroborated by a variety of evidentiary avenues that do not themselves include reliance upon the substantive theory of interest, in which case we properly think of ourselves as mainly testing the substantive theory "in reliance upon" the previously corroborated auxiliary. But there is no necessity for this state of affairs, and even in the physical sciences, both the substantive theory and the auxiliary assumptions may have the same status at a given time with regard to the prediction of a particular observational result. Thus, for example, a conjecture concerning the structure of crystals and a conjecture concerning X-rays were both problematic at the time the epoch-making series of experiments on the use of crystals in X-ray diffraction was initiated. It is well known in the history of physics that these experiments *simultaneously* corroborated conjectures concerning the inner structure and molecular distances in crystals as well as the wave nature of X-rays and the order of magnitude of X-ray wavelengths. It would have been

impossible to answer the (seemingly reasonable) question, "Which are you presupposing in testing—the X-ray theory or the crystal theory?" The answer to that question, which physicists were too sophisticated to ask, would be *"Both,* conjointly and simultaneously."

An extremely simple, clear, unproblematic example from the physical sciences, of estimating an unobserved true value via two fallible measures in the absence of a "direct, independent" avenue to either's accuracy, is Rutherford and Geiger's calculation of total (observed *and unobserved*) alpha-particle scintillation events via the numerical relations among the one-observer and both-observer tallies. High-school algebra cancels out the two nuisance parameters of "observer efficiency" (Segré 1980 pp. 103-104).

9. When a substantive theory and the auxiliary assumptions are both somewhat problematic at a given stage of our knowledge, a misprediction of observational fact refutes the conjunction of the two, and does not ordinarily give us much help in deciding which of them—perhaps both—must be modified or abandoned. It is of great importance, however, to recognize the other side of this coin, to wit, that successful numerical prediction, *including convergent numerical predictions by different epistemic avenues*, mediated by the conjunction of the two conjectures, tends to corroborate their conjunction—both of them receiving evidentiary support thereby.

I have spent some time on this matter of assumptions and conjectures because in developing the estimation methods, I shall rely on what would ordinarily be called "assumptions" that readers may find doubtfully plausible, as I do myself. If these "assumptions" are perceived (wrongly) as *necessary premises* for the use of the convergent methods I propose, the latter will of course appear to lack merit. If on the other hand these "assumptions" are seen as part of the system of interconnected conjectures or hypotheses, which network of interrelated hypotheses gives rise to a prediction that certain experimentally independent (instrumentally nonredundant) estimates of an unknown, unobserved quantity will agree tolerably well, then a different evidentiary situation is presented to the critic. Instead of being told, "If you are willing to assume A, then method M_1 will estimate something, and so will method M_2, and so will method M_3 …," we say, "If we conjecture the conjoined auxiliary and theory (A.T), then nonredundant methods M_1, M_2, M_3, M_4, should lead to consistent numerical results." If those concordant numerical results would be, without the conjunctive conjecture, "a strange coincidence," so that the predicted agreement of the methods takes a high risk, goes out on a limb, subjects itself to grave danger of refutation—then, as Popper says, surmounting that risk functions as an empirical corroborator. The usual use of "assumption" means something that is not itself testable. Whereas in my "assumptions," made to get the theoretical statistics going, I intend that the consequences of those assumptions are to be tested by the agreement of the empirically independent statistical estimates. So I shall deliberately avoid the use of the word "assumption" in favor of the words "hypothesis" and "conjecture," with the adjective "auxiliary" when appropriate.

Another way of viewing this matter is via the concept of *consistency tests.* (See Meehl 1965, 1968, 1978, 1979; Golden & Meehl 1978; and Meehl & Golden 1982.) If the parameter we desire to estimate is not directly measurable, either because its physical realization occurred in the past or because it is an inherently unobservable theoretical entity, we cannot subject the instrument of measurement, whether it is a physical device or a statistical procedure for manipulating observational records so as to obtain a theoretical number, to direct validation—the ideal procedure when available. Thus, for

example, in assigning evidentiary weight to certain signs and symptoms of organic disease in medicine, we now usually have available fairly definitive results, reported by the pathologist and the bacteriologist, concerning the tissue condition and its etiology that underlay the presented symptoms and course of the disease. But before knowing the pathology and etiology of the disease, the only way the physician could draw conclusions as to the differential diagnostic weight of the various symptoms and complaints, course of illness, reaction to treatments, and the like was by some kind of coherency, the clinical "going-togetherness" of the findings. Most of the theoretical entities inferred in the behavioral sciences, such as human traits and factors, or even temperamental variables of genetic origin, have to be "bootstrapped" in this way (Cronbach & Meehl 1955; Golden & Meehl 1978, 1980; Meehl & Golden 1982).

There is nothing inherently objectionable about such procedures of bootstrapping by statistical manipulation of correlations between what are presumed to be fallible observable indicators of the desired unobserved latent state of affairs; but it is also known that many, perhaps most, methods of factoring and clustering such fallible phenotypic indicators of an alleged latent causal-theoretical entity are subject to an undue amount of ad hockery, to such an extent that some cynics in the social and biological sciences reject all such methods of factor and cluster analysis on the grounds that they will lead us to find factors, clusters, types, and the like even when there aren't any there (cf. Meehl 1979). It has therefore become an important task of the statistician and methodologist to derive procedures that will constitute internal checks on the validity of the attempted bootstrapping. One cannot say in advance that he will always succeed in finding a factor or a cluster if it's there, and that the method will surely give the right diagnostic weights to the fallible observable indicators; but some assurance should be provided. Similarly, an investigator should be reasonably confident that he will not wrongly find a factor, cluster, entity, or syndrome when it's not really there. We don't merely want statistical gimmicks that will "bring order out of disorder." We want procedures that will discern order only if it is latently present. We want, as Plato said, to carve nature at its joints.

In some of my own work developing new taxometric statistics for studying the genetics of mental diseases like schizophrenia, I have devoted attention to deriving multiple consistency tests by means of which the validity of a given inference made from a pattern of fallible observables to, say, the presence of the dominant schizogene would be possible. My reasoning again involves rejecting the usual "assumption-versus-hypothesis" model of inference in favor of a joint conjecture model. We avoid writing the conventional A: $(f_1 \cdot f_2 \cdot f_3) \rightarrow T$, which reads, "*If* you are willing to *assume* A, then the three facts $f_1f_2f_3$ *prove* the theory T." (This formulation is objectionable anyway because, as the logicians point out to us, facts in science never, strictly speaking, "prove" theories but only test them.) Instead we write the neo-Popperian $(A.T) \rightarrow (f_1 \cdot f_2 \supset f_3)$, which we read as, "Conjecturing the conjoined auxiliary and substantive theories A and T, we see this conjunction entails that facts f_1 and f_2 imply fact f_3." Further, since fact f_3 has a sufficiently low antecedent probability on $(f_1 \cdot f_2)$ lacking the theory, such a prediction takes a high risk, that is, exposes itself to grave danger of being falsified. If it escapes falsification, the conjoint conjecture (A.T) is corroborated.

The normal way, and the way that will be taken in this paper, of conceptualizing such low prior probability facts that are capable of subjecting the theory to a high risk of refutation is that the predicted facts are numerical point values or relatively narrow

ranges. So we don't really rely on the "assumption" A in justifying our numerical methods, the traditional way of putting it which I am here strenuously opposing. Rather, we rely on the unlikely coincidence of converging numerical values in order to corroborate the "assumption" together with the main theory of interest that combines with it to generate an otherwise improbable numerical consequence.

Classic examples of this kind of thing abound in the exact sciences such as physics, chemistry, and astronomy. For instance, there are over a dozen independent ways of estimating Avogadro's number. In the early stages of the development of physical chemistry, each of these ways could have been criticized as relying on unproved "assumptions," and some of them were only doubtfully plausible at the time. But whatever the initial plausibility of the "assumptions relied on" (in one way of looking at each method singly), these are nonredundant independent methods and they converge at the same estimate of the number of molecules N in a mole of a substance. The fact of their numerical convergence tends to corroborate the correctness or near correctness of the assumptions (and, in some cases, the robustness of the method under departures from the idealization) *and,* be it noted, *at the same time gives us a high degree of scientific confidence in the theoretical number*, *Avogadro's constant*, *thereby inferred* (cf. Perrin 1910). The inference is made by multiple paths, and even though each path "relies on" what taken by itself might be only a moderately plausible assumption (and from now on I shall, I repeat, avoid this term in favor of "auxiliary hypothesis" or "conjecture"), once the convergence of these paths is found as an empirical fact, we view the a priori shakiness of the "assumptions" in a different light.

I do not, of course, intend to browbeat the reader by arguments from the philosophy of science into accepting the merits of my suggested procedures. I merely use them as motives of credibility in anticipation of the first criticism that springs to mind as soon as one begins to examine each of these methods, namely, that it "relies on assumptions" that are, to say the least, problematic and that some critics would think very doubtful.

Before commencing exposition of the methods, I remind the reader of our general strategy at this point, so as not to bore him with irksome repetitions of the preceding methodological guidelines. Each method will involve mathematical idealizations "based on" strong (and, hence, antecedently improbable) "assumptions" about the paleontological situation—both the *objective situation* (the "state of nature") and the *investigative situation* (the process of fossil discovery). I shall simply state these strong assumptions, emphasizing that they are intended as auxiliary conjectures whose verisimilitude (Popper 1962, pp. 228-237; 1972, pp. 52-60) can only be guessed at from the armchair in our present state of knowledge, and, therefore should mainly be tested by their fruits. The fruits, as indicated above, consist of the agreement or disagreement among the four essentially independent (nonredundant) methods of estimation.

The quantity we are interested in may be called the *completeness coefficient.* For a given form, like a species, that coefficient, of course, takes on only the two values 1 or 0; a fossil of the species is known to us, or it is not. A certain kind of animal lived on the earth, but is now extinct, and we may have no knowledge of it because we have not found even a single recognizable fossil. For larger groups, such as families, orders, classes, or phyla, the completeness coefficient is not confined to the two-valued "found at least once (= 1)" or "never found, even once (= 0)," but may refer to the proportion of different species belonging to the larger group that have been found at least once. Example: If

there were in the history of organic life on the earth 200 species of ungulates that are now extinct, and at the time of computing the completeness index, paleontological research had discovered 70 of these species, the ungulate completeness index would be .35. Then in the eyes of Omniscient Jones, the paleontologist knows of the existence of only about one in three of the species of ungulates that have existed but are now extinct. (The exclusion of extant species from the reference class that defines the denominator of the completeness index will be explained later, when I present Dewar's Method of Extant Forms.) The basic unit tallied in forming this index I shall take as a species, although arguably, because of the perennial disagreements between "lumpers" and "splitters" in taxonomy, it might be preferable to take as basic unit the genus, as was done by Dewar and Levett-Yeats (cited in Dewar & Shelton 1947, pp. 61-63, 71). I do not see that anything except reliability hinges upon this choice, so I shall refer to species. The question of how high up in the hierarchical taxonomic tree the larger reference group should be chosen is also arbitrary to some extent, but not completely so because we have two countervailing methodological considerations from the standpoint of computing good statistics. Since the coherency of the four methods is, in the last stage of my procedure, to be examined by correlating estimated total numbers of extinct species within a given taxon, it is desirable to keep the larger classification sufficiently small so that there can be a goodly number of taxon-numbers to correlate. Thus we would not wish to take as denominator the number of phyla because then the correlation coefficients computed for estimates based on the four methods would be based on an *N* of only 20-30. On the other hand, for reasons of stability of the individual completeness coefficients computed by each method, we do not want to be dealing with very small numbers of species within each larger category. In the case of vertebrates, for instance, something like number of carnivora, classified at the level of the order, might be a reasonable value. As in the species/genus decision, nothing hinges critically upon this one. It's a matter of convenience and of statistical stability. So, in each of the methods, I shall refer hereafter to *species* as the basic taxon unit tallied, and have arbitrarily selected *order* for the larger group whose species tally generates a proportion.

Method I: Discovery Asymptote

The first completeness estimator I call the *Discovery Asymptote Method.* It relies on the simple intuitive notion, widely used in the life sciences, that if an ongoing process of producing or destroying some quantity is known to have a physical limit, and if the process goes on in a more or less orderly fashion, one can fit a curve to data representing the stage reached at successive points in time; and then, from that fitted curve one can arrive at an estimate of the total quantity that would be produced (or destroyed) in the limit, i.e., in infinite time. In our fossil-finding situation, the total (unknown) quantity is the number of species of a given extinct order that would ever be found if paleontological search continued forever, and the ordinate of such a graph at year 1983 represents the number of species of the order that are known to us today. For such a method to have any accuracy for the purpose of estimating the total number of extinct species of an order, the asymptote of total cumulative finds at infinite time must be approximately identifiable with the total that have lived. So the first idealization "assumed" (better, *conjectured*) is that every species in the order has been fossilized at least once and that fossil preserved somewhere in the earth's crust. The qualitative, common-sense considerations for and

against that idealization have been set out above. At this stage we conjecture that the huge number of individual organisms of a species suffices to counteract the very low probability of an individual specimen being fossilized and preserved. We do *not* treat the exhaustion of the earth's crust by paleontological digging as an idealization (as we did with the previous conjecture), because that idealization is mathematically represented by the asymptote, which the mathematics itself tells us we shall never reach. But the abstract notion is, counterfactually, that paleontologists continue to explore the earth "forever," or at least until the sun burns out, and that no cubic yard of earth, even that under the Empire State Building or at the bottom of the sea, remains permanently unexamined. That this is of course a technological absurdity does not distress us, because we do not require to reach the asymptote physically but *only to estimate where it is.*

As to the choice of the function to be fitted, there is, as in all such blind, curve-fitting problems not mediated by a strong theory, a certain arbitrariness. But familiar considerations about growth processes can help us out. We know we do not want to fit a function that increases without limit, however slowly, as we get far out on the time scale, because we know that there is an upper bound, the number of different species of animals that have lived on the earth being, however large, surely finite. A choice among curves would be based upon trying them out, fitting the parameters optimally to each candidate function, and selecting the "best fit" by examining the residual sums of squares when each candidate function has been optimally fitted by least squares or other appropriate method. The most plausible function is what is sometimes called in biology and psychology a "simple positive growth function." Let N be the total number of species of, say, the order Carnivora that have ever lived. On the idealized assumption that the record in the earth (not, of course, the record in our museums or catalogues) is quasi-complete—that is, that every species of Carnivora is represented by at least one fossil somewhere in the earth's crust—the number of species known to us at time t, say in the year 1983, is the ordinate y in the known record of Carnivora. Then the residual $(N - y)$ is the number of species yet remaining to be found. Adopting the notion of a constant search pressure, say S (which, of course, is a gross oversimplification over the whole history of paleontological search, but I shall correct that in a moment), we postulate that the instantaneous rate of change of the ordinate is proportional to the number of species yet remaining to be discovered, the constant of proportionality being the search pressure S. This gives us the familiar differential equation for simple growth processes:

$$[1] \qquad \frac{dy}{dt} = \mathrm{S}(N - y) \ .$$

Integrating this differential equation, and determining the constant of integration by setting $y = 0$ at $t = 0$, we obtain the integrated form

$$[2] \qquad y = N(1 - e^{-\mathrm{S}t}) \ ,$$

which is a familiar expression to many biological and behavioral scientists, sometimes being called the "exponential growth function," sometimes "simple positive growth function" because it often approximates empirically the growth of organisms and their parts, as well as the growth of certain states and dispositions. For example, the psychologist Clark L. Hull had a predilection for fitting exponential growth functions in describing mammalian organisms' acquisition of habit strengths (Hull 1943). The growth

function attained visibility in the life sciences partly from the frequent use of it, or its autocatalytic form, by the Scotch-American-Australian biochemist T. Brailsford Robertson, who was highly regarded and widely known for his work on growth. Behind a trial-and-error attempt to fit empirical data, and successfully applied to a wide array of growth processes in the biological and social sciences, was the underlying concept from physical chemistry that the velocity of a monomolecular reaction is at any time *t* directly proportional to the amount of chemical change still remaining to occur. But aside from that historical scientific "rationale," the fundamental notion is our conjecture, in relation to any process of change, that the momentary velocity of the process is proportional to the amount of change still remaining to occur. So that for a fixed search intensity or find pressure, represented by the postulated constant S in the exponent of the growth function, the momentary rate of finding new species will, on an essentially random model, depend upon how many species there are left to find.

Experience in the life sciences has shown that the vast majority of data we get in the empirical world tend to follow a very few curve types. Among those most frequently encountered are the straight line, the parabola, the exponential curves of growth and decay, the logarithmic function, the Gompertz curve, the normal probability curve, and its integral, the normal ogive. Empirical curve fitting is still considered in applied mathematics to be partly a common-sense, intuitive, or theory-based business. Nevertheless we can, in principle, always select between two competing "admissible" curves on the basis of which fits better in the sense of least squares or another defensible criterion (e.g., Pearson's method of moments) when each competitor curve's parameters have been optimally fitted. In the case before us, the justification for giving special attention to the exponential growth function as a prime competitor is theoretical, for the reasons stated. Some functions, such as a straight line, a parabola, or an exponential function can be excluded confidently on theoretical grounds because they do not approach an asymptote. The total number of extinct species "to be found" (even on the fantastic assumption that paleontologists will dig up the whole earth's crust before the sun burns out) is finite, and that simple physical truth puts a constraint on our choice of functions.

But there is a pretty unrealistic idealization involved in the above text preceding our differential equation, namely, that the search intensity S remains the same from whatever arbitrary time dates (the year 1800 or 1700, say, or whatever one chooses) we begin to plot data for the science of paleontology at a search intensity sufficient to be significant in curve fitting. One must surely suppose that the search intensity has considerably increased from that time to the present, if for no other reason than the great increase in the number of professional paleontologists and the financial support provided for their work.

Thus the parameter S cannot have a fixed value over the last two centuries or so of paleontological searching, but must itself be increasing more or less steadily with time. As a better approximation, we could try, in the integrated form of the equation, to write instead of the parameter S a variable search intensity $s = a + bt$. This is. of course, still an approximation based upon the idealization that rate of finding is a linear function of amount of searching, and that amount of searching is a linear function of time. The first of these has considerable plausibility as probably a not-too-serious idealization, so that the mathematical expression of it would be robust under slight departures; but the second one, that the search pressure is a linear function of time, is presumably not a close approximation to the truth. Putting nonlinear functions into the exponent complicates the

curve fitting process considerably, however, and it also leads to greater instability in assigning numerical values to the parameters. If one wanted to get some realistic notion of just how grossly distorted the linear growth of a search intensity exponent is, he had better go not to the historical record of fossil finds but to a more direct (social) measure of the search intensity. An alternate variable—like number of professionally active paleontologists in the world at a given time—would presumably serve as an adequate proxy; but we could also plot number of expeditions recorded, number of research grants given, or even number of dollars spent. The intent there would not be, of course, to determine the parameters, which cannot be done in this fashion, but to determine the exponent's likely function form so as to ascertain whether an approximation to s as a linear function of t is too grossly out of line with facts. These are crucial details that cannot be explored profitably here. If the linear model for variable search intensity exponent s as a function of time were provisionally accepted as good enough for the present coarse purpose, the integrated form of the growth function would then become

$$[3] \qquad y = N(1-e^{-(a+bt)t}) = N(1-e^{-(at+bt^2)}) \ .$$

So we have a more complicated growth function with the same parameter asymptote (total fossil species N to be discovered in the limit), but the growth constant is now replaced by a second-degree polynomial as a function of time.

If it is found that this modified exponential growth function with a variable growth rate graduates the empirical data for a given order, say, Carnivora, satisfactorily (which means in the practice of curve fitting that it graduates it about as well as the leading competitor functions that have some theoretical plausibility, and preferably somewhat better than these others), then the best fit by least squares determines the parameters N, b, and a. If the original cruder form were an adequate approximation to our paleontological discovery data, it determines the two parameters N and S. The parameter N is the asymptote that the search process is approaching in the limit of infinite time exhausting the earth's crust. On our idealization that the record held in the rocks—*although not our record at any time*—is quasi-complete, this asymptote is an estimator of the total number of species of the order we are investigating that ever lived and have become extinct. At the present moment in time (year 1983) the value of y in the fitted function is the number of species of Carnivora that have thus far been discovered in the fossil record. So the completeness index, the "percent adequacy" of the Carnivora fossil record, is simply y/N, that is, the proportion of all (known + unknown) extinct species of Carnivora that are known to us as of 1983. The same curve-fitting operation is independently carried out for each of the orders of animals with hard parts that we have decided to include in our study of the agreement of the four methods.

The ratio y/N is of considerable intrinsic interest, in that most paleontologists would be astonished if it were greater than 50 percent, indicating that we know today the majority of species of extinct Carnivora; so that the record would be much more complete than it is customary to assume in paleontological writings. But the number we want for our subsequent correlational purposes when investigating the "methods" consistency is not the percentage 100 (y/N) but the absolute asymptote N. The statistical reason for this will be explained later. The point is that we have fitted the discovery growth function for the order Carnivora and have recorded its asymptote N, the total estimated species of Carnivora, *both those known to us and those not presently known*, for our later use. We

do the same thing for primates, rodents, and so forth, each time the curve-fitting process terminating in an estimation of the unknown parameter *N*, the total number of species of extinct members of that order, known and unknown.

The Discovery Asymptote Method just described relies on a process taking place through recorded historical time, and each extinct species (of a given order for which the cumulative discovery graph is plotted) appears only once as a datum contributing to the determination of that curve: i.e.. it appears associated with the abscissa value of *t* at its first discovery. This is important because it means that the *number* (absolute frequency) of individual fossil specimens of the same extinct species plays no role whatever in determining the parameters of the function fitted in the Discovery Asymptote Method. The second method I have devised pays no attention to date of find, or otherwise to an ongoing time-dependent discovery process, but is wholly "cross-sectional." It ignores time but attends instead to the very thing that the first method systematically excludes from consideration, namely, how many individual fossil specimens of each known extinct species are available at the present time. This is crucial because in order for two or more methods of estimating an unknown quantity by means of some proxy measure (such as we must rely on in paleontology) to be mutually corroborative, the methods must not be redundant. That is, they must not be essentially numerically trivial reformulations of the same raw data, or of summary statistics of those raw data which are known algebraically or empirically, via dependence on some powerful "nuisance variable" (Meehl 1970), to be highly correlated.

Method II. Binomial Parameters

The second method, which I call the Method of Binominal Parameters, again takes off from a well-known intuitive idea in the application of statistics to the life sciences. This idea is that the distribution of the frequencies with which an improbable event is known to have occurred is a reflection of its improbability and can therefore, under suitable circumstances, be employed to estimate that underlying unknown probability. I shall deliberately sidestep the deep and still agitated question of the basic nature of the probability concept, which remains unsettled despite the best efforts of mathematicians, statisticians, and philosophers of science. The reader may translate "probability" either as a partially interpreted formal concept, the notion apparently preferred by most mathematicians and by many statisticians who are not primarily identified with a particular substantive science; or as some variant of the frequency theory, still very popular despite recent criticisms; or, finally, as a propensity—the view favored by Sir Karl Popper. I hope I am correct in thinking that these differences in the basic conceptual metaphysics of the probability concept, surely one of the most difficult notions to which the mind of man has ever addressed itself and, unfortunately, one of the most pervasively necessary in the sciences, do not prejudice the argument that follows.

We are forced to get our method off the ground by postulating a strong idealization that we are confident is literally false, and whose verisimilitude is unknown to us. Here again the verisimilitude is not estimable directly, and for that reason is only indirectly testable by the consistency approach, i.e., by its convergence with other nonredundant modes of estimating the record's completeness. Consider a particular order (say, Carnivora) and a particular extinct species (say, the saber-tooth tiger) belonging to that order. We represent the resultant at any time (e.g., 1983) of the paleontological discovery

process as a collection of individual, physically localized excavations, each of which we arbitrarily designate as a "dig." Since this idealization of a dig is unavoidably imprecise and coarse, we consciously set aside difficult questions about its area, returns to excavate nearby, distinguishing between individual investigators or teams, and the like. We also set aside the question of whether multiple specimens of the same species found in the same dig are counted as one or many, a decision that should be made on the basis of the availability of the relevant data in the catalogued museum record. Choice among these first-level data reductions (instance classifications and counts) may, without vicious circularity, be made on the basis of their comparative conformity to the conjectured mathematical model. We ask: "Which is more orderly (= Poisson-like)—frequency of finds as tallied by method A or by method B?" If, for example, counting three saber-tooth skulls as one find because they were found by one team digging in a "small area" reveals operation of a cleaner Poisson process than would counting them as three, then the former tallying rule is the rational choice.

Then we consider the abstract question of whether in a given dig, a fossil specimen of the particular species is or is not found. (We are not here tallying actual digs, of course—only conceptualizing them.) We then conceptualize the probability p of its being found. From the frequency interpretation of the probability concept, what this amounts to is that we conceptualize the huge class of digs in the hundreds of thousands (setting aside the complication "what the paleontologist is looking for") and consider what proportion of them contain a saber-tooth tiger fossil.

Let us designate as n the total number of digs on which this propensity (or relative frequency) of finding is based. Given the usual idealizing assumption of essential independence (not in this case too unrealistic), the probability of finding a saber-tooth in all n digs is p^n; the probability of finding a saber-tooth in all but one of the digs is $np^{n-1}q$ (where $q = 1 - p$ is the "not-find" probability); and so on with the familiar series of terms in the expansion of the binomial $(p + q)^n$. Then we have as the last term of this sequence of probabilities, each of which specifies the probability of finding precisely k saber-tooth fossils among n digs, the probability q^n, which is the probability of failing to find a single saber-tooth fossil among any of the n digs. This epistemologically regrettable event would correspond to our total ignorance that such an organism as the saber-tooth tiger ever existed. Consequently the last term of the binomial expansion is the one of interest for the present purpose, its complement $(1 - q^n)$ being the completeness coefficient when we apply it over the entire order Carnivora.

On the further idealizing assumption that the same find probability over digs applies to each species of the order being investigated, or on the weaker assumption that the mean value of these probabilities can be applied without too much distortion over the board (a matter of robustness), let the total number N of extinct Carnivores be multiplied by the expansion of $(p + q)^n$. If each probability term in the expansion were thus multiplied by the unknown species number N for the order Carnivora, this would generate the theoretically expected *absolute frequencies* with which the various species of the order are represented as fossils in the museum catalog. Of course, we do not know the number N, and the point of the procedure is to estimate it. What the theoretical model just developed tells us is that the observed frequencies of species represented by only a single fossil, by two fossils, by three fossils, and so on up to hundreds or thousands of fossils (as we have of the woolly mammoth or the trilobites), is latently generated by the process of

multiplying the terms of the binomial by the unknown constant *N*.

The big fact about the *empirical* graph we obtain by plotting (starting at the left) the numbers of extinct species of the order Carnivora that are represented by only one fossil, by two fossils, by three … —a tallying process performable empirically without having recourse to the above theoretical model of how these frequencies were generated—is of course the fact that the frequency of species corresponding to the extreme left-hand value of the abscissa variable: "number of fossils representing the species" is an empty cell. It represents the number of extinct species of Carnivora for which the paleontologist has at present found not a single fossil. Of course that number is not known, because those are the species we do not at present know to have existed at all. The basic idea of the Binomial Parameters Method is that we can make an extrapolative estimate of the unknown number of species in that last column by studying the properties of the distribution of those in the remaining columns.

Before presenting the development of the procedure for this extrapolation, which unavoidably involves some alternatives, each of which must be tried on real data, I shall attempt to give the reader an intuitive appreciation of the way of thinking underlying this method by beginning with the very coarse-grained question, "Is the record at least half complete?" That is, taking the order Carnivora as our example, can we at least conclude, without attempting more precise parametric inferences about the underlying binomial properties, whether we presently have in the museums at least one fossil representative of over half of the extinct species of Carnivores?

Suppose we conjecture the contradictory, with all contemporary evolutionists, that the record is considerably *less* than 50 percent complete. That is, of all the species of Carnivora that ever walked the earth, paleontology in 1983 knows of fewer than half. For over half of them, not a single fossil has been found. This means that the latent quantity $q^n > 1/2$. Since on our idealized model the find probabilities are conjectured to be approximately equal and independent for different carnivores, it doesn't matter whether we consider the terms of the binomial here or the result of multiplying the expanded binomial by the unknown total Carnivora species number *N*, since these are proportional. We can therefore consider the observed distribution that will run from the extreme right, where we have hundreds or even thousands of fossils of certain carnivores, and move to the left with diminishing numbers. In the far left region we have a group of species, each of which species is represented by only two fossils; then to the left of that is abscissa value = 1, which designates those species each of which is known to us by virtue of only a single fossil specimen; and then finally the empty cell of interest, the abscissa value = 0, corresponding to those species of Carnivora that are unknown to us because no specimen has yet been found as a fossil even once. In the graph of the distribution of finds, this unknown, were it plotted, would be the highest ordinate. But if $q^n > 1/2$, the sum of all the other terms in the binomial, i.e., $(1 - q^n)$, must be $< 1/2$. Therefore each individual term except q^n must be less than 1/2. In particular, the term adjacent to the unknown term at $x = 0$, i.e., the probability corresponding to the number of species that are represented only by a single fossil in the record known to us, must be less than the height of the unknown ordinate at $x = 0$. Since it is a theorem of algebra that the binomial—even if extremely skew and whatever its parameters—cannot have multiple local modes, it follows that there cannot be another mode anywhere to the right of the mode latently existing for $x = 0$. But if the coefficients ever increased as we moved to the right, there

would be two local modes and the theorem would be violated. Hence the model requires that there cannot be an increase as we move to the right, toward the species columns represented by large numbers of individual fossils. Since no two terms can be equal, it follows that as we move to the right toward species represented by very large numbers of fossils, the graph is everywhere decreasing. So we see that if the record is less than 50 percent complete, the observed distribution of numbers of species represented by various numbers of individual fossil finds must slope everywhere upward toward the left. This corresponds to a latent situation known in statistics as a "J-curve " found in studying low probability events, such as radioactive decay, behavioral studies of social conformity, and the like.

So it is interesting to note that before starting a more complicated mathematical treatment of the data by studying the statistical properties such as skewness and kurtosis in the higher moments of the distribution form, choosing between a binomial and a Poisson distribution for this purpose, etc., we can already see that a universally-held doctrine of paleontology, to the effect that the record is considerably less than half complete, can be subjected to rather direct test over a group of orders by simply plotting the "find replication" numbers for various species. The test prediction is that these graphs, which should be fairly stable statistically (the numbers involved being large provided that the unit tallied is genus or species and the supra ordinate taxon is broad enough, such as order or class), should reveal a monotone function rising to the left. Further, on the assumption that the completeness index is less than 50 percent and usually taken to be very considerably less than this (the vast majority of species in any extinct order being unknown to us), we must infer that $q^n >> 1/2$. Then, since the complementary fraction $(1 - q^n) << 1/2$ must be dispersed over a very large number of integral values on the abscissa (covering the range of species of fossil replication numbers up to those fossils represented in the hundreds or thousands), it follows that we expect a nearly flat and very low graph until we get to the extreme left end, where it is expected to undergo a steep rise shortly before we come to the empty interval at $x = 0$. More will be said below about this problem of shape in a discussion of the binomial and Poisson distribution question.

Since the physical situation of fossil discovery reflected in empirical statistics is surely one in which the individual event of interest ("a find"—of a single specimen) has extremely low probability, this being countervailed to some (unknown) degree by the sizable number of paleontological digs, a highly skew binomial distribution can be anticipated with confidence. The highly skew binomial distribution is customarily represented by the Poisson distribution, the terms of which involve only the one parameter $m = np$. where n is large and p is small, their product being a (proper or improper) fraction near to one. It is known from wide experience in a variety of scientific contexts (e.g., emission of alpha rays, occurrence of industrial accidents, overloading of telephone exchanges, rare cell counts in a haemocytometer) that the Poisson distribution provides an excellent fit to observational data arising from this combination of a small probability for an event with a large number of opportunities for it to occur. Mathematically the Poisson distribution is generated from the binomial by a limiting process in which, holding the product $np = m$ fixed, we imagine p going to 0 as n increases without limit, and the defined binomial moments are then examined. So the Poisson distribution is, strictly speaking, an approximation to the precise numerical values we would get by using the binomial, even for cases of extreme asymmetry in its complementary probabilities. (This explains why I

have called Method II "Binomial Parameters," to designate the general case—covering the the unlikely situation of only slight skewness—and the rough bound-setting process described supra. Mainly I rely in what follows on the special properties of the Poisson distribution for numerical estimates, so Method II could as well have been labelled "Poisson Parameter" instead.) One should, however, beware of assuming that we must take *very* extreme values (infinitesimally small p's and gigantic n's) to get a good Poisson approximation, since even moderate degrees of asymmetry are graduated quite well. (Cf. Kyburg 1969. pp. 146-151, noting that Table 6.1 has a misprint heading penultimate column; Lewis 1960, pp. 243-258; Feller 1957, pp. 135-154; Hays 1973, pp. 202-208; Cramér 1946, pp. 203-207; Yule & Kendall 1940. pp. 187-191.) Fitting a Poisson distribution with the single parameter m to data that would be well graduated by the binomial for asymmetrical ($p << q$) would give substantially the same answer, an answer that would differ numerically from the precise one by amounts less than the errors introduced by our raw data and the literal falsity of the conjectured discovery model (independence, crude approximations in defining what is "a dig" and "a find," and the like). The mathematical theory of statistics provides no criterion for when the Poisson distribution approximates the asymmetrical binomial closely enough, since this is not a mathematical but a "statistical tolerance" question having no precise mathematical meaning. Appendix I suggests a way of using the binomial proper instead of the Poisson approximation to it, but it has the disadvantage of estimating more latent parameters via expressions of unknown robustness under departures from the idealized model.

As explained above, the basic problem is an extrapolation problem (not soluble by any simple extrapolation procedure that we can trust), namely, to fill in the frequency of species occupying the extreme left-hand interval in the distribution of number of species represented by various numbers of single specimens from 0, 1, 2, 3, … to very large numbers of specimens. No observation is available for the number of unknown species whose find frequencies are 0, i.e., we don't know how many species there are of extinct Carnivora for which we have not found a single fossil. By definition, since we haven't found any of them, we don't know how many there are. The Poisson distribution conjecture, suggested by the plausible physical considerations supra (small find probability but numerous digs), engenders two submethods for estimating the missing ordinate at $k = 0$.

The basic idea of the first Poisson method is that if the distribution, including the unknown ordinate of the graph at $k = 0$, is generated as theoretically conjectured by a Poisson process, we can make use of a simple, well-known relation between the successive terms of a Poisson distribution to solve for the missing species frequency. In the graph, the observations plotted are the frequencies with which different species of Carnivora are represented in the record by 1, 2, 3, **…,** k individual fossil specimens. So each ordinate of the graph is $Np(k)$ where N is the fixed (unknown) number of species of Carnivora that ever lived. Since the N will cancel out in the first step, we neglect it for present purposes. Terms of the Poisson distribution giving probabilities of k and $(k + 1)$ fossil specimen finds are

$$[4] \qquad p(k) = e^{-m}\frac{m^k}{k!}$$

$$[5] \qquad p(k+1) = e^{-m}\frac{m^{k+1}}{(k+1)!}$$

so the ratio of two successive terms is

$$[6] \qquad \frac{p(k)}{p(k+1)} = \frac{m^k}{k!}\cdot\frac{(k+1)!}{m^{k+1}} = \frac{k+1}{m}$$

Taking logarithms and their differences (the Δ's in standard interpolation theory) we obtain

$$[7] \qquad \log\frac{p(k)}{p(k+1)} = \log p(k) - \log p(k+1) = \Delta^1$$

and taking the second difference $\Delta^2 = \Delta^1_i - \Delta^1_j$,

$$[8] \qquad \log\left(\frac{k+1}{m}\right) - \log\left(\frac{k+2}{m}\right) = \log\frac{k+1}{k+2} = \Delta^2$$

That is, the second order differences Δ^2 of the log frequencies are a simple sequence of fractions, the unknown parameters having cancelled out. At the rare-find end of the distribution these fractions are small integers, their generative law being the addition of one [= 1] to both numerator and denominator, thus:

$$[9] \qquad \begin{aligned} &\textit{for} \quad k = 4, \quad \Delta^2 = 5/6 \\ &\qquad\quad k = 3, \quad \Delta^2 = 4/5 \\ &\qquad\quad k = 2, \quad \Delta^2 = 3/4 \\ &\qquad\quad k = 1, \quad \Delta^2 = 2/3 \\ &\qquad\quad k = 0, \quad \Delta^2 = 1/2 \end{aligned}$$

Note that we can test the observed frequencies for their closeness to the Poisson before proceeding to our main step, which is extrapolating to the unknown frequency N_k at $k = 0$, "species for which no fossil has been found."

Taking the Δ's at this low-frequency end (k = 2, 1, 0) with the variable being *log raw observed frequency*, we have

$$[10] \qquad [\log N(2) - \log N(1)] - [\log N(1) - \log N(0)] = \log (1/2),$$

so the unknown is obtained:

$$[11] \qquad \log N(0) = 2 \log N(1) - \log N(2) + \log (1/2),$$

the desired unobserved frequency on our graph at $k = 0$.

A second submethod of employing the Poisson distribution properties to estimate the unknown missing species frequency relies upon the fortunate mathematical fact that the first and second moments of a Poisson distribution are equal (as is also the third moment, not used here but perhaps usable as a check on numerical consistency of the results). The frequencies of the Poisson distribution being probabilities for discrete values of k, the whole system is, of course, multiplied by the unknown total number of species of Carnivora N to generate the observed raw frequencies for numbers of fossil representatives per species. Let N^* be the total observed frequency for the known fossil species, that is,

[12] $N^* = N(1) + N(2) + \ldots + N(k) + \ldots$

and unstarred N is the true total number of species, known and unknown,

[13] $N = N^* + N(0),$

the second term on the right being the missing point on our "finds" graph at $k = 0$. Take the mean of known points

[14] $$\overline{x} = \frac{1}{N^*}\sum^{N^*} x$$

The true unknown mean is

[15] $$= \frac{1}{N}\sum^{N} x = \frac{1}{N}\left(\sum^{N^*} x + \sum^{N(0)} x_0\right)$$
$$= \frac{1}{N}\sum^{N^*} x + 0$$

i.e., the unknown cell, if known, would affect the N we divide by but not the sum of x's, since this sum is a count of "number of specimens found" and the missing cell at $k = 0$ refers to those species *not* found. Then since

[16] $$\mu = \frac{1}{N}\sum^{N^*} x$$

$$\sum^{N^*} x = N\mu$$

and

[17] $$\sum^{N^*} x = N^*\,\overline{x}$$

[18] $$N\mu = N^*\,\overline{x}$$

hence, as $N = N^* + N(0),$ $\left(N^* + N\left(0\right)\right)\mu = N^*\,\overline{x}$

so

[19] $$\mu = \overline{x}\frac{N^*}{N^* = N(0)} = \overline{x}\frac{N^*}{N}$$

which is in two unknowns μ and N, so indeterminate as yet. But we now consider the second moment

$$\mu_2 = \frac{1}{N}\sum^{N} x^2 - \mu^2$$

[20] $$= \frac{1}{N}\left(\sum^{N^*} x^2 + \sum^{N(0)} x_0^2\right) - \mu^2$$

$$= \frac{1}{N}\left(\sum^{N^*} x^2 + 0\right) - \mu^2$$

Designate by S^* the observed second moment (about zero) for intervals $k = 1, 2, 3, \ldots$

[21] $$S^* = \sum^{N^*} x^2$$

in the above. Then we have the true mean and true second moment

$$[22] \qquad \mu = \overline{x}\frac{N^*}{N}$$

$$[23] \qquad \mu_2 = \frac{1}{N}S^* - \mu^2$$

The "completeness index" for total species count is

$$[24] \qquad C = \frac{N^*}{N}$$

i.e., the proportion of all species, known and unknown, that are known to us.

If the underlying find-function is Poisson, we may infer that the first and second moments are equal, writing

$$[25] \qquad \mu_1 = \mu_2$$

which from the preceding amounts to saying that

$$[26] \qquad \overline{x}\frac{N^*}{N} = \frac{1}{N}S^* - \mu^2$$

$$= \frac{1}{N}S^* - \overline{x}^2\left(\frac{N^*}{N}\right)$$

and transposing

$$[27] \qquad \left(\frac{N^*}{N}\right)^2\overline{x}^2 + \left(\frac{N^*}{N}\right)\overline{x} - \frac{1}{N}S^* = 0$$

which in "completeness index" notation reads

$$[28] \qquad C^2\overline{x}^2 + C\overline{x} - \frac{1}{N}S^* = 0$$

and dividing by completeness index

$$[29] \qquad C\overline{x}^2 + \overline{x} - \frac{1}{N^*}S^* = 0$$

which transposed and dividing by $\overline{x}^2$ yields

$$[30] \qquad C = -\frac{1}{\overline{x}} + \frac{1}{N^*\overline{x}^2}S^*$$

$$= \frac{S^* - N^*\overline{x}}{N^*\overline{x}^2}$$

$$[31] \qquad = \frac{\sum^{N^*} x^2 - N^*\overline{x}}{N^*\overline{x}^2}$$

and all quantities on the right are observable from the fossil frequency distribution of knowns ($k > 0$), so we obtain the completeness index C. Dividing C into N^* gives us our

estimate of the total number of Carnivores, found and unfound.

I have no grounds for preferring one of these methods to the other, and the existence of two Poisson-based methods helps us to get a consistency check. It might be argued that since they both rely on the conjectured latent Poisson process, they *must* be equivalent; and hence their numerical agreement is tautologous and cannot bear on anything empirical. That is a mistake, because the algebra does not permit a direct derivation of the equality relied on in the second method from the equality relied on in the first method. (I invite the reader to try it if he thinks otherwise.) We have here a nice little point in philosophy of science (one sometimes forgotten by scientists in making claims of equivalence between *consequences* of theories) that two theorems may flow from the same postulates—and hence each provides a test of the postulates empirically and an estimate of quantities referred to in the postulates theoretically—but the two theorems do not follow *from one another.* It's really a trivial result of the difference between one- and two-way derivation chains in a formalism. There is a "deep" sense in which they're equivalent, in that they both spring from the same conjecture. But since they are not mutually derivable from the formalism alone, they tend to reinforce one another and corroborate the conjectured latent process that leads to each of them. The nice distinction to see here is that between a truth "*of* the formalism" (i.e., an algebraic identity) and a conjecture concerning the physical world formulable "*in* the formalism," which then has a consequence *derivable via the formalism.*

It is perhaps unnecessary to say that one can test the observed frequencies for their closeness of the Poisson distribution before proceeding to either of these main steps of extrapolating to the unknown value at $k = 0$, "species that have *no* fossils found as yet." Here again a significance test is not appropriate. We know it doesn't exactly fit. The question should be, "Is it a bad fit or a reasonably good one?"

If the two methods agree within a reasonable tolerance and if no systematic bias can be shown theoretically to obtain, nor is there other reason why one of them should be privileged, one might simply strike an average between the two estimates of the missing frequency in the class interval $k = 0$ for species not found. Then, as in the Discovery Asymptote Method, one records the estimated total number of species $N^* + N(0) = N$ for the order Carnivora. The same process is carried out for each of the orders being considered, thus generating an array of estimates of total extinct species over the various orders. These are elements in a second column of estimates for the frequency of extinct species (known and unknown) for the various orders, corresponding to estimates previously obtained by the Discovery Asymptote Method.

Method III. The Sandwich Method

The third method, which I have christened "Sandwich" for lack of a less clumsy term and for reasons that will be obvious, is superficially the most straightforward of the three methods invented by me, and at first blush seems as simple as the fourth method (Douglas Dewar); but a little critical reflection shows the simplicity to be illusory. The method turns out to be the most complicated and treacherous of the four, and despite many hours of manipulating algebra and calculus, Monte Carlo trials, and consultation with four mathematically competent colleagues, I am not entirely satisfied with the results. Intuitions on this one can be quite misleading, but the core idea seems (to me and everyone to whom I have explained it) pretty clearly sound. Rather than wait upon

publication of this paper for a more satisfactory estimating procedure, or for a simpler approach via the same basic idea (which my intuitions persist in telling me does exist, though I have not been able to hit upon it), I present it here for criticism and improvement.

The basic idea of the Sandwich Method is simple and compelling, despite the complications that appear when we seek an exact formula. If we know that the saber-tooth tiger existed at a certain point in geologic time, and we know that he was still extant at some subsequent point of geologic time, then we know that there must have been saber-tooth tigers at all times during the intervening period, since the process of evolution does not exactly duplicate itself by the formation of a species identical with one that has previously become extinct (Dollo's Law). On this assumption, an evolutionary truism not disputed by anybody, one approach to the question, "How complete is the fossil record for extinct species belonging to the order Carnivora?" runs as follows: We presuppose a high confidence determination of the geologic time for the earliest stratum in which a reliably identified saber-tooth tiger fossil has been found, and similarly the identification of other fossil(s) of the saber-tooth tiger in a reliably dated stratum different from the first one. (We pay no attention in the Sandwich Method to *when* it was found nor to *how many* individual fossil specimens are found—so the Sandwich Method is nonredundant with the first two methods.) We then consider strata dated anywhere in between these two anchor points in geological time, and inquire whether at least one saber-tooth fossil has been found in any of them. Since we know that the species must have existed for the entire intervening time period between the two extreme values anchored, the fossil record is incomplete, with regard to this extinct species, for the time interval between the anchoring fossil finds if we have not found it anywhere in that interval.

Using such an approach, we may consider every extinct species of Carnivora for which anchoring fossils have been reliably identified in strata accurately dated, and separated by some specified geologic time interval, as defining a "sandwich." The two strata anchoring the species in geologic time constitute, so to speak, the "bread slices" of the sandwich, outside of which no fossil specimen of this species has been found. The completeness index is then based upon tallying, for the various species sandwiches thus defined, what percentage of such sandwich-associated species are also represented, following the metaphor, in the "innards" of the sandwich. This tally would be made by inquiring, for each species defining a sandwich by bread slices, whether at least one fossil specimen is found in a stratum reliably dated as lying between the two bread slices. We then calculate from this tally of all extinct species that define a sandwich (so as to be available for such a calculation), what proportion of them are also represented in the innards of the sandwich (so to speak, in the Swiss cheese or ham as well as in the defining bread slices). The resulting number would be a crude measure of how complete the record for Carnivora is. One may conceive loosely of a "total antifind pressure," the combined factors opposing the [forming + preserving + discovering] of fossils, a quantitative property for extinct species collectively of a given broad taxon. We try to estimate the net impact of these counterfind pressures on the taxon Carnivora by tallying representations of species in that taxon for geologic periods during which each is known to have existed, and the numerical value of that impact is then used to infer how many "unfound" Carnivorous species there were.

As I say, the reasoning seems initially quite straightforward: If we know that a

species existed during a certain time period, which we can know by knowing that it existed before that time period (in the bottom bread slice of the sandwich) and after that time period (because we find it in the top bread slice), then a failure to find even a single fossil specimen of it anywhere in any strata dated between the slices of the geological sandwich thus defined tallies one "incompleteness instance."

But the percentage completeness thus computed as a crude fraction of total sandwich-defining fossil species is not a valid estimate of true "completeness of the record," as we shall see shortly. I do not say that such a crude completeness index, obtainable by summing such tallies over all extinct species of Carnivora that define various sandwiches of different sizes (and different absolute locations in the time scale), is utterly without value. For example, it could set a safe lower bound that might be illuminating in our present weak state of knowledge about completeness. But if we desire to obtain a quantitative estimate of the incompleteness index that would yield a number for missing species that could be meaningfully correlated over orders with the (allegedly correspondent) numbers of missing species as estimated from the Discovery Asymptote Method and the Binomial Parameter Method, we need a Sandwich Method that will not suffer from a gross systematic bias.

It is easily seen that a crude count of within-sandwich representations over the set of sandwich-defining species suffers from an unknown but non-negligible amount of bias downward. That is, the completeness of the record is seriously underestimated by this crude, unweighted procedure. Two or three home-baked Monte Carlo trials (with differing find-probability assignments and modest species $N = 100$) can be done with a desk calculator, and the bias is so large as to persuade immediately. The first reason for this is that in calculating the percentage completeness in this crude way by simply summing the unweighted tallies of species that define sandwiches and that are also found at least once within the sandwich, we must, of course, ignore the two (or more) fossils definitive of the sandwich slices since, obviously, the proportion represented here = 1; otherwise we wouldn't know about the species in the first place, or have a sandwich available to make the count upon. For sandwiches in which the defining bread slices are far apart in geologic time, so that the sandwich is very "thick" in its innards (a geological "Dagwood"), we have many geologic time slices in between, where there exists an opportunity for a saber-tooth tiger fossil to be found. But consider species which in reality (unknown to us) had a very short geologic life span, i.e., in which the true species longevity amounts to only three time intervals, so that the fossil specimen found in the latest interval and the one in the earliest interval defining the two bread slices of the sandwich leave only one layer of Swiss cheese in between. If that inner sandwich does not contain a specimen, this species is tallied as a failure of completeness. But of course we have in fact found him twice, in two distinct time slices. We need not have an accurate algebraic expression for the size of this bias to realize that short longevity species will, for any small or even intermediate level of *within-sandwich probability of a find,* be contributing adversely to the completeness index. More generally, the completeness index is being unavoidably reduced by the neglect of the fossils found in the bread slices that define the sandwich for each species because, of course, if we haven't got the bread slices, we haven't found a species.

A second consideration is that it would be somewhat surprising, under usual evolutionary assumptions, if there were no correlation whatever between the within-sandwich-

slice probability p_i and the true species longevity L_i. On the average, at least if we consider animals falling within a given broad taxon such as the order Carnivora, one would expect variables such as *total number of specimens alive at a given time* and their *range of geographical distribution* to be correlated with ultimate species longevity, and this should generate a correlation between longevity L_i and find-probability p_i within slices of the sandwiches.

I shall present two methods that conjecture (and test for) a relatively fixed p_i over the L_i's, then a third method that approximates variable p_i's. Finally, a complete, nonapproximative method, but one that involves a rather messy system of nonlinear equations, will be explained. Appendix II presents some interesting theorems concerning the latent generating system that are not used in the four methods but should help to fill out the picture and perhaps inspire readers more mathematically competent than I to derive other estimators or consistency tests motivated by the "sandwich" concept.

Our idealized body of observational data consists of geological sandwiches defined by fossils of the various extinct species of our exemplary Order Carnivora. Units of geologic time (rather than named strata corresponding to times) will be expressed for convenience as a tenth portion of the time corresponding to the species of greatest longevity as indicated by the thickest sandwich. Although some species will, under our circumstances of incomplete knowledge, have defined sandwiches of shorter duration than the species' true longevity L, we take the longest longevity as being safely estimated by the thickest sandwich found for any species of Carnivora. This is a reasonable assumption because, despite the fact that some species of great longevity may have been found in considerably smaller sandwiches, it is improbable that *none* of the species of the extremest longevity has defined a sandwich of that thickness. Since even one species defining a sandwich that matches its own longevity suffices to determine this upper value, for us to be in error in the downward direction would require that *every* species of the maximum true longevity fails to be found at its sandwich edges—a highly unlikely situation. (Should this happen, we have a somewhat smaller maximum longevity than the true—not a vitiating error for the method.) Dividing the longest species survival time into ten equal time units, I shall designate each subinterval as a *decichron*. Then the true unknown species longevities range from $L = 1$ through $L = 10$ decichrons, and it will be convenient to refer to species longevities with subscripts as L_{10}, L_9, …, L_1. To avoid repetition of the language "true" or "actual but unknown" longevity, hereafter the term "longevity" will be taken to refer, as will the capital letter L, with or without subscripts, to the true historical time period through which the species persisted.

The sandwiches defined by identifying one or more fossils of a species in two distinct geological time intervals separated by another, this being the necessary condition to have a sandwich for which presence or absence of the species fossils within the sandwich contents can be tallied, will be called a *total sandwich* when we include the two sandwich-defining outer layers (metaphorically, the "bread slices"). The phrase *inner sandwich* will refer to the layers between the two bread slices. So that if s = the total sandwich including the bread slices, and k = the number of layers within the sandwich of size s, we have for every sandwich size, $s = k + 2$. The term "sandwich," when unqualified, means one having innards $k > 0$, $s > 2$.

It will not be confusing that sometimes I use "degenerate sandwich" to designate cases $s = 2$ and $s = 1$, two adjacent bread slices (no innards), or a single bread slice.

Although we are interested ultimately in species of longevities down to L_2 and L_1, they will not enter our initial calculations because such short-lived species cannot (except through errors of identification which we are here setting aside) generate a sandwich of the minimum possible sandwich size $s = 3 = k + 2$ where $k = 1$. Neglecting these short-lived species does not introduce a bias because the equations first employed refer only to those of longevity greater than $L = 2$.

For all species of a group that have defined sandwiches of any given size s, we could ask whether or not that species is represented in any given intra-sandwich slice. "Represented" here means found *at least once* as a fossil in that slice, as here we are not concerned with how many (as we were in the Binomial Parameter Method) nor with when found (as in the Discovery Asymptote Method). The tally of representations per slice is a crude initial index of "completeness" for a slice, but it is of course not the number we are ultimately interested in. I shall refer to that number as the *crude observed slice-p*, when it is averaged over the k within-sandwich slices for sandwiches of total thickness $(k + 2)$. These slice probabilities for sandwiches of a given size, as well as their dispersion over the slices, constitute another portion of our observational data. The relationship between the dispersion of the slice probabilities and the average slice probability of a sandwich of a given size should be related, on a substantially random model of within-sandwich finds, by a simple formula associated with the nineteenth-century satistician Lexis (Uspensky 1937, pp. 212-216). I do not reject that approach, as I think it should be pursued further; but it is not the one I adopt presently. The slice-p value will instead be treated as an inferred "latent construct" variable. The latent state of affairs includes the distribution of longevities $p(L_{10})$, $p(L_9)$, …, $p(L_3)$; and to each of these longevities, which include, of course, species that we have not found at all (even in a single bread slice incapable of defining a sandwich, i.e., species not known to us), there is a corresponding conditional probability of generating a sandwich of a certain size. Also there is a probability, for a species that has generated a sandwich of that size given a true longevity L, of its being represented anywhere within a specified inner sandwich slice and therefore of its being represented in the sandwich at all. If p is the single-slice probability, its complement $(1 - p) = q$ is the nonrepresentation-slice probability. Then q^k is the probability of not being represented anywhere within the sandwich, so its complement $(1 - q^k)$ is the probability of its being represented in a sandwich *if that sandwich is defined*.

Designate species longevities by L_1, L_2, …, L_{10} and the proportions of carnivorous species having these longevities as $p(L_1)$, $p(L_2)$, …, $p(L_{10})$. These proportions would appear as Bayes's theorem "prior probabilities" if we could estimate them. The slice-probability that a species of longevity L_1 is found represented in any one time slice $\Delta T = (1/10)L$ (L = largest longevity = 10 decichrons) is designated by p_1, the subscript indexing the source longevity. The conditional probability that a species of longevity L_1 generates a sandwich of total ("outside") thickness s is designated $p(s/L_1)$, which is the same as the probability of a sandwich whose "innards" thickness (excluding the bread slices in which the sandwich-defining fossils were found), is $s - 2 = k$.

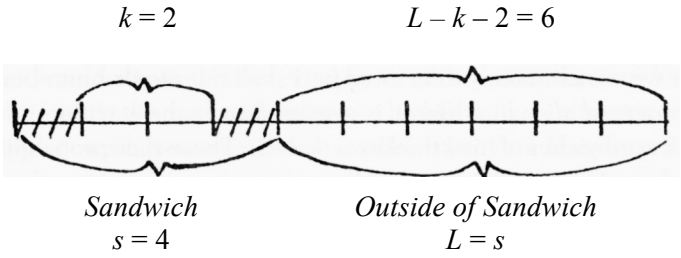

Figure 1

What is the probability that a species with longevity L generates a sandwich of inner size k? The diagram illustrates with $L = 10$, $k = 2$ how the slice-probability p_i determines this sandwich-generating conditional. The longevity L_{10} is divided into 10 equal decichrons. One way to get $k = 2$ from $L = 10$ is shown, the earliest bread slice (shaded interval) occupying the extreme left position. The slice-probability p gives p^2 as probability of the joint event "finding a fossil in these two (specified) bread slices." The sandwich occupies $s = k + 2 = 4$ decichrons, leaving $L - k - 2 = 6$ slices "outside the sandwich." All these slices must be empty, this joint event having the probability q^6 (no fossil representation in any of them, $q = 1 - p$ being the complement of the slice-probability p). Since there are 6 additional positions it could occupy, the composite event can occur in 7 ways, counting the one shown. The general expression for a sandwich of inner thickness k arising from longevity L is therefore

$$[32] \qquad p(k / L) = (L - k - 1)p^2 q^{L-k-2}$$

We shall employ this conditional probability of sandwich size on a generating longevity in deriving the second and third approximate methods now to be explained.

I present first an approximation that may seem rather crude, but that turns out to be remarkably close, as the reader may easily verify by trying a few selected numerical examples representing various combinations of parametric values. I myself am satisfied that it is sufficiently accurate, given our modest goal—to get rough estimates of the fossil record's completeness. Next I present two improvements on this approximation. The precisifying development I present finally will probably not be much better than these approximations, since the precision expressed in the formalism will not be physically realized. Departures from the idealization expressed in the more "exact" formalism will, I fear, wash out any theoretical improvement in exactitude over the rather crude methods I now explain.

If the slice-q values associated with each longevity were known to us, we could compute the true incompleteness index by multiplying the base-rate probability $p(L_i)$ for a species of longevity L_i by the probability that it is *not* found represented in any of the time slices from 1 to i. But this (assuming independence) is simply the slice-q improbability raised to the L_i power. The sum of these products over all longevities is then the true incompleteness for all species, all longevities, in the order Carnivora, thus:

$$[32] \qquad I = p(L_{10})q^{10} + p(L_9)q^9 + ... + p(L_1)q^1$$

We do not know the numerical values of any of these 20 quantities, but we are now going to approximate them crudely. We may first inquire whether there is evidence that the q's depend strongly on the L's. Since we do not know the L's, we take as proxy the

associated k's, i.e., the observed sandwich thicknesses. There is a strong correlation between the L's and the k's because a short longevity cannot be associated with a sandwich larger than it, although a large L may generate a few very small sandwiches. Hence if there were an appreciable correlation between the longevities L_i and their associated slice-q values in the underlying stochastic generating situation, there would be a corresponding (although somewhat attenuated) correlation between observed sandwich thicknesses k and the q-values that generate the intra-sandwich completenesses.

Given a slice-q value q, the probability of a species not being found in any of the k slices of a sandwich of inner thickness k (on a random model of fossilization and finding) is simply q^k. If only a single value of q were the underlying generator of the observed incompleteness tally for sandwiches of size k, that generating q could be obtained simply by taking the kth root of the observed incompleteness index I_k for sandwiches of that size. So what we do first is compute the observed proportions of species not represented in the sandwiches of various sizes against the sizes $k = 1, 2, \ldots 8$, to see whether there is evidence of a trend line. I do not think it appropriate to carry out formal significance tests; but if desired, that can be done. Mainly what we want to know is whether there is a moderate to strong relation between incompleteness as an estimator of the q values and the sandwich sizes taken as proxies for the true generating longevities. If there is not—if the graph is flat (or even if it bounces around somewhat but shows no clear trend)—we choose the kth root of the crude incompleteness as our guess at the underlying slice-q value for all longevities.

Then the grand true incompleteness index over all longevities would be obtained by multiplying the longevity proportions by the corresponding powers of q and summing as in the above equation [33]. This we cannot do, not knowing the longevity distribution, but we again make a crude approximation by ignoring the distribution of longevities—treating it in effect as if it were rectangular—and raising the kth root of the grand crude incompleteness index I_c to the power L/k. Here k is the mean observed sandwich size that we compute directly from the data. L is unknown to us, but assuming approximate symmetry of the distribution of longevities (whether they are normally, platykurtically, or rectangularly distributed), since they range from a low of $L = 1$ to a high of $L = 10$—the mean of these integers for a symmetrical case being 5.5—this value is taken as our approximation to the unknown L. In effect what we are doing is substituting an exponent that is a ratio of two estimated mean values of roots and powers, and applying this fractional exponent to a crude sum that is the unweighted, uncorrected incompleteness index over all sandwich sizes. It's essentially a matter of raising a sum to a certain power rather than summing the component values taken to differing powers initially before adding them. Oddly enough, one finds by taking various combinations (including the cases of equal or extremely variable q-values and the cases of rectangular or normally distributed longevities) that these coarse approximations result in a systematic error of only one or two points in the second decimal place in estimating the true underlying incompleteness index I. This works because of the sorts of numbers we are dealing with, to wit, fractional variables raised to powers that are ratios of small integers.

Perhaps a better approximation, although it is not obviously better given the probable distortions of the idealized model and the statistical instability of the components being added, is obtainable by solving for the distribution of longevities as an intermediate step. If the initial graphing of the quantities $(I_k)^{1/k}$ for various k's reassures us that the gener-

ating q-values do not correlate strongly with the longevities, we may use the conditional probability of a longevity generating a sandwich of a certain size (empty or occupied, not relevant) derived in the text above, equation [32]. Then the expected value of the proportion of sandwiches of a given size can be written as a sum of products of the unknown longevity probabilities by the appropriate conditionals for each longevity's generating a sandwich of each size k, including the two degenerate cases $s = 2$, $s = 1$. The sum of these products is the expected value of sandwiches of size s over all longevities, the entire system of such equations being soluble since the conditional probabilities can be computed from the quasi-constant q-value we had estimated over all sandwich sizes. The proportion of sandwiches of each size being found directly from the data, we can solve for the unknown latent longevity probabilities $p(L_{10})$, $p(L_9)$, … $p(L_3)$, $p(L_2)$, $p(L_1)$. When these longevity probabilities are thus estimated, we then multiply each of them by the constant value q raised to its appropriate exponent L_i for that longevity, and sum to get the grand true incompleteness.

A somewhat better approximation makes use of the fact that a species defining a sandwich of total (outer) size s must have a longevity $L \geq s$. This being so, we can restrict the range of latent longevity sources L_i that underlie the observed incompleteness index for a sandwich of inner size k. This "averages" the unknown latent generating values (L_i, L_j …) over a smaller range, so that striking a plausible representative value (I use midpoint) involves less "smearing" than taking kth root of total crude incompleteness I_c over all sandwich sizes. Furthermore the procedure locates enough points so that if a straight line seems to be a reasonable graduation, we shall wash out some of the fluctuations by fitting that line. The procedure is as follows:

Consider the observed sandwiches of size $s = 7$ (innards size $k = 5$). These sandwiches can have arisen only from longevities $L = 7$, 8, 9, or 10. Although the latent slice-p values vary over these four longevities, the composite crude incompleteness index $I_c(5)$ is employed as a "smeared" value in the approximation. We estimate a representtative latent slice-q underlying the observed sandwiches of size s by taking the kth root of the crude observed incompleteness $I_c(k)$ for sandwiches of that size. We do not know the latent longevity distribution, and the relationship between that and the smearing effect is rather complicated. (Thus if extremes of longevity are somewhat rarer, the distribution not being strictly rectangular at least at the ends, this softening influence at the ends is modified by variation in the shape of the conditional probability graphs of $p(s_i/L_i)$ for various sandwich sizes on each longevity.) What we do is strike a middle value for the possible longevities capable of originating sandwiches $s = 7$, taking the midpoint of longevities $L = 10$, 9, 8, 7, that is, $L_m = 8.5$. (We are not bothered by the fractional longevity since, of course, the original time division was quite arbitrary, simply taking tenths of the maximum longevity for a given taxon of intermediate level.) We do the same for sandwiches of all eight "complete" sizes, i.e., those for which an incompleteness index is computable because it has innards $k \geq 1$. We fit a straight line to these eight points (notice that one of them, $s = 10$, does not involve any smearing approximation). The slice-q values for the two short longevities ($L = 1$ or 2) that can give rise only to degenerate sandwiches ($s = 1$ or 2) are then read off the extrapolated line.

Given these 10 estimates of the latent slice-q's, and taking each one's complement $p_i = (1 - q_i)$, we can then write directly the expected values of observed frequencies of sandwiches of all 10 sizes as functions of the conditional probabilities $p(s_i/L_i)$. We have a

system of equations linear in the unknown latent longevity frequencies $N(L_i)$ thus:

$$
\begin{aligned}
N(s_{10}) &= N(L_{10})\, p(s_{10} / L_{10}) \\
N(s_9) &= N(L_{10})\, p(s_9 / L_{10}) + N(L_9)\, p(s_9 / L_9) \\
N(s_8) &= N(L_{10})\, p(s_8 / L_{10}) + N(L_9)\, p(s_8 / L_9) + N(L_8)\, p(s_8 / L_8) \\
&\vdots \qquad\qquad \vdots \\
N(s_1) &= N(L_{10})\, p(s_1 / L_{10}) + N(L_9)\, p(s_1 / L_9) + \ldots + N(L_1)\, p(s_1 / L_1)
\end{aligned}
\tag{34}
$$

Solving these for the latent longevity frequencies $N(L_1)$ by putting our *observed* sandwich frequencies $N(s_j)$ on the left, we add $N(L_1) + N(L_2) + \ldots + N(L_{10})$ to obtain the estimated total number of species of all longevities

$$
N_t = \sum_{j=1}^{10} N(L_j) \tag{35}
$$

Numerical trials and inspection of the graphs of conditional probabilities show that this approximation is remarkably good, despite the smearing of the latent values. It is likely to do about as well as the "exact" procedure described next, because the exact procedure involves the solution of a system of 16 equations of high degree in which both the slice-*p* values and the longevity distribution frequencies are unknowns, whereas in this third approximation, the conditional probabilities are first obtained by substituting the slice-*q* values in the formula so that the system of equations to be solved is linear.

The only "exact" solution I have managed to come up with makes use of the fact that we possess two items of information about all 8 of the nondegenerate sandwich sizes ($k \geq 1$, $s \geq 3$), namely, for each such sandwich size we know its empirical frequency N_k and its crude completeness index I_k . Since these observable quantities are expressible in terms of the latent longevity-frequencies and the latent slice-*q* values, we obtain a soluble system of 16 equations in 16 unknowns. The remaining four latent values associated with short longevities (N_2, N_1, q_2, q_1) we get by extrapolating and check for consistency by predicting the two degenerate sandwich frequencies $N(s_2)$, $N(s_1)$. The "exact" method, using all the sandwich frequency and incompleteness data at once, proceeds as follows:

Let us confine attention first to the nondegenerate sandwiches ($s = k + 2 \geq 3$). It is convenient to employ raw species frequencies rather than probabilities, avoiding the normalization problem to $\Sigma P_i = 1$. Only species with longevities $L_i \geq 3$ can generate these sandwiches, so we have to consider eight longevities and eight (associated) slice-*p*'s. It is convenient to treat the complementary slice-improbabilities q_i instead of the p_i's themselves. The latent situation is then

Longevities:	$L_{10}, L_9, L_8 \ldots L_3$
Base-frequencies:	$N_{10}, N_9, N_8 \ldots N_3$
Slice-q's:	$q_{10}, q_9, q_8 \ldots q_3$

All the *N*'s and *q*'s are unknowns, and there is no constraining equation on either of them (except, of course, that $\Sigma q_i < 8$, which is not helpful).

The observed sandwich frequencies are expressible in terms of these latent quantities (writing "observed" on the left, although strictly speaking these equations yield only the *expected values* of the several species *N*'s). Thus:

$$
\begin{aligned}
N(s_{10}) &= N_{10}\, p(s_{10} / L_{10}) \\
N(s_9) &= N_{10}\, p(s_9 / L_{10}) + N_9\, p(s_9 / L_9) \\
&\vdots \qquad\qquad\qquad \vdots \\
N(s_3) &= N_{10}\, p(s_3 / L_{10}) + N_9\, p(s_3 / L_9) + \ldots + N_3\, p(s_3 / L_3)
\end{aligned}
\tag{36}
$$

For each sandwich (inner) size k there is available an observed crude incompleteness index I_k. These observed indexes for each inner thickness k are compounded from latent incompleteness indexes arising from all sources with a sufficient longevity ($L \geq k + 2$) to generate such a sandwich; but for each such latent source L_i the expected incompleteness among the subset of sandwiches *it* generates will be q_i^{k}. These latent component incompletenesses have expected contributions to the sandwiches in proportion to the proportional composition of each sandwich size's source species. Thus if $\mathrm{p}(L_{\mathrm{i}} / k_i)$ is the posterior ("inverse," Bayes' Theorem) probability that a sandwich of size k_i has originated from a source of longevity L_i, we can write the expected values of observed sandwich incompletenesses as follows:

$$
\begin{aligned}
I(k_8) &= p\left(L_{10} / k_8\right)\, q_{10}^8 \\
I(k_7) &= p\left(L_{10} / k_7\right)\, q_{10}^7 + p\left(L_9 / k_7\right)\, q_9^7 \\
&\vdots \\
I(k_1) &= p\left(L_{10} / k_1\right)\, q_{10}^1 + p\left(L_9 / k_1\right)\, q_9^1 + \ldots + p\left(L_3 / k_3\right)\, q_3^1
\end{aligned}
\tag{37}
$$

This system of 8 equations adds no new unknowns, because the inverse probabilities $p(L_i / k_j)$ are Bayes' Theorem functions of the priors and conditionals in the system [36]. Hence we have a total system of 16 equations in 16 unknowns that we can solve.

After solving these to obtain the 8 latent longevity-frequencies $N(L_{10})$, $N(L_9)$, …, $N(L_3)$ and their 8 associated slice-improbabilities $q(L_{10})$, $q(L_9)$, …, $q(L_3)$, we still have to estimate the latent species frequencies $N(L_2)$, $N(L_1)$ for species too short-lived to produce a nondegenerate sandwich, and the slice-improbabilities q_2, q_1 associated with them. This we do by extrapolation, thus: Plot the $N(L_i)$ values against the longevities (L_{10}), (L_9), …, (L_3) and fit a curve—a straight line should be good enough, tested for linearity if desired—to the 8 points now known. The two N's at the low longevity end are then assigned by reading the ordinates off that fitted line. Similarly we obtain q_2, q_1 by reading off a line fitted to the eight pairs $(L_i,\ q_i)$ for $L_i \geq 3$. A consistency check for the trustworthiness of these extrapolated values is available by writing the equations for the 10 "sandwich" frequencies (now including the degenerate cases $s = 2$, $s = 1$) in terms of the latent generating values.

We now possess numerical estimates of the values of all 20 latent variables conjectured to generate the observed distribution of sandwich sizes and the crude incompleteness indexes tallied for nondegenerate sandwiches. The longevity frequencies $N(L_{10})$, $N(L_9)$, …, $N(L_1)$ can be summed to obtain the total (found and unfound) carnivorous species number, and the completeness index is the ratio of the known species tally to this sum:

$$
C = N^* \Big/ \sum_{i=1}^{10} N(L_i) \tag{38}
$$

We can also obtain this estimate from Equation [33] supra (where $I = I - C$) and the two should agree. We record this species number in the third column of our master table of four estimates, in the row for the order Carnivora.

IV. Method of Extant Forms (Dewar's Method)

The fourth and final method of completeness estimation was not invented by me but by Douglas Dewar. I am especially indebted to his rarely mentioned work because it was my reading of the exchange of letters between Dewar and H. S. Shelton that first interested me (alas, over thirty years ago, when I set the problem aside) in the possibility of estimating the completeness of the fossil record, although I was informed by colleagues in zoology and geology that such a numerical estimate is obviously "impossible." It has been easy to minimize the importance of Dewar's contribution because despite having some professional distinction as an ornithologist and, as his writing clearly shows, an extraordinarily intelligent and resourceful mind, he was also a die-hard anti-evolutionist, whose espousal of creationism was apparently not associated with conservative Christian theology. A scientifically legitimate objection to the Method of Extant Forms is that it relies (in the traditional non-Popperian notion of "rely" to mean "justification") on questionable "assumptions" concerning some quantitative equivalences that are, to say the least, doubtful. In the eyes of an evolutionist who wishes to dispute even the possibility of a completeness index for the fossil record, these equivalences would not be acceptable.

The essential notion of Dewar's Method is extremely simple, and if the questionable assumption about certain equivalences were granted, it is the most straightforward and least problematic of the four methods here discussed. In all three of the methods I have devised, we are concerned, so to speak, in estimating the denominator of a fraction—the completeness index—from its numerator, given some kind of conjectured function relating the two. The numerator is the number of extinct species of a group known to us from paleontological research, and the desired denominator is the unknown total number of species of that group (such as all animals with hard parts, or all vertebrates, or all carnivores, or all ungulates) that have ever walked the earth. We are moving from the known to the unknown—not inherently a methodologically sinful procedure, as Popper points out—and the confidence we have in that epistemic movement hinges at least partly on the plausibility or independent testability of the alleged functional relationship between the known term in the numerator and the unknown term in the denominator. I say only "at least partly," because the essential independence or nonredundancy of the three paths for estimating an unknown number provides a strong Popperian test of the accuracy of the estimation procedures, as I said in my introductory remarks. The relative importance of (a) *the a priori plausibility* of the postulated functional relationship between the known and unknown quantities and (b) *the mutual corroboration provided by numerical agreement*, no one presently knows how to assess, particularly if the agreement permits large error tolerances. But as we desire here to estimate an unknown quantity of great scientific importance, we are willing to attempt it even if the percentage of error for each estimate is quite large. (I shall say more below about assessing the convergence of values.) But Dewar's method, the Method of Extant Forms (as I may call it to indicate its character, as the other three were labelled, although I think Douglas Dewar should receive posthumous recognition for his method), is unlike my three methods, because in Dewar's Method the desired denominator is known to a high

accuracy for many taxa. That is due to the fact that the tally of species in the denominator consists of extant forms. The discovery rate for new extant species of the order Carnivora is now extremely small and has been for many years, so that zoologists can say that contemporary science "knows about" the existence of close to 100 percent of the species of carnivores living today. Or, moving up from order to even such a broad taxon as class, of "large mammals" only two new species have been discovered in the last 150 years (the giant panda and the okapi). So, to answer the first-level question, "What is the completeness of the fossil record for *extant* species of Carnivora?" we have a simple and straightforward computation, to wit, we divide one number known with high accuracy (how many species of extant carnivores have been found at least once as a fossil) by another number known with very high accuracy (how many extant species of Carnivora there are). No complicated algebra or doubtful auxiliary postulates about randomness, linearity, etc., are required for applying Dewar's Method. For those groups of animals in which the current discovery rate of new species is considerable, even as adjudged by taxonomists of a "lumper" rather than a "splitter" bias (such as the hundreds of thousands of insect species that are still being added to at a respectable rate by entomologists), the method presents more serious problems, so those taxa might better be set aside for that reason. The vexed question of which phyla, classes, orders, and so forth should be candidates for the four completeness index estimation procedures here presented, I discuss below.

But corresponding to the delightful simplicity of Dewar's Method is, of course, the transition from a completeness index for extant forms to the one we are really interested in, i.e., the completeness index for extinct forms. We pay for the extreme methodological straightforwardness and algebraic simplicity of Dewar's index, the mathematics of which is so much simpler and the embedding justificatory text so much more plausible, with a slippery and dangerous assumption, that *the completeness index for extant and extinct forms is about the same size fraction.* With my neo-Popperian approach, i.e., auxiliary conjectures as part of the network to be tested by the fact of convergence rather than pure faith "assumptions" (which an optimist may be willing to rely on and a pessimist will reject), I don't face quite the problem that Dewar had with Shelton, when the latter objected to the use of extant forms as a basis for estimating the completeness of the fossil record for extinct forms. The reason is again that the numerical convergence of nonre–dundant estimation procedures is taken as a Popperian test of the validity of each, including its auxiliary conjectures, however plausible or unplausible these may seem when examined qualitatively.

Dewar's Method does not strike me as so implausible as Shelton seemed to think, though. One must remember that we are not comparing groups of very diverse kinds, where the number of the individual specimens, the geological period during which they flourished, of generations per decichron, the evolutionary rate, the geographical distribution, the ecological niche, and so forth, are often very different (as they would be if one were comparing the completeness index for extant bears with the completeness index for, say, extinct echinoderms). What we are comparing is the completeness index for animals like the polar bear with animals like the saber-tooth tiger, because the final stage of our procedure is to involve the correlation of estimated species numbers over various groups, and these groups are defined as animals of roughly the same sort, i.e., carnivores, ungulates, crustaceans, or whatever. So long as the conditions of the earth are such that a given order, say, carnivorous mammals, can exist in appreciable numbers at all (part of

this, of course, is simply Lyell's uniformitarian assumption that evolutionary theory postulates instead of catastrophism), it is not obvious that the conditions operating statistically to determine the laying down of a fossil, its preservation, and contemporary searchers' luck in finding it are very markedly different between extant and extinct species *of the same order or family.*

The idealized conjecture relied upon in Dewar's Method is that the terrestrial conditions for preservation of the hard parts of an animal as a fossil; the distribution, intensity, and duration of fossil-destroying geologic occurrences such as erosion and faulting; and—what presumably will not have varied appreciably—the statistical odds of the paleontologists search locating one and identifying it, have remained, in their aggregate influence, roughly comparable for extant carnivores and extinct carnivores. As suggested above, we conceptualize (without claiming to quantify causal-theoretically) a sort of "anti-find pressure," the net influence of these collective factors acting adversely to paleontological discovery. The incompleteness index I_k (= species not found/total species) is some monotone function of this antifind pressure, and the rationale of Dewar's Method is the conjecture that the antifind pressure is likely to be about the same for extant and extinct species of the same broad taxon, as carnivores, primates, ungulates, etc. The manner in which we employ the simple completeness index based on fossil finds of extant species is also simple. The fraction C^* obtained on extant forms is taken as an estimator of the corresponding completeness index C for extinct forms of the same kind. e.g., order Carnivora. We therefore take the number of species N^* (= number of extinct carnivorous species found at least once as a fossil), and, writing $N^* = NC$, we solve for N, the desired total number of extinct species of Carnivora, known and unknown.

Reflection on Dewar's Method quickly suggests to the mind a different and powerful use of the fossil data on extant forms, namely, a direct test of the numerical accuracy of what I may call "Meehl's Methods." The main use of Dewar's Method is as an estimator of the completeness index for extinct forms, where as we have seen, it has advantages over my three methods in its conceptual and mathematical simplicity, but with the attendant disadvantage that it relies on the vague concept of "aggregate counter-find pressures," postulated to be approximately the same for extinct and extant forms of the same intermediate level taxon (e.g., order Carnivora) to rationalize the extrapolative use of the index to the extinct groups. Assigning some degree of antecedent probability to that auxiliary conjecture, we test its approximate validity indirectly, by means of the empirical test of convergence among the four methods taken together, as explained in the next section below. Although this indirect test is perfectly good scientific procedure, whether we think in a neo-Popperian fashion or simply reflect on the history of the developed sciences in their use of auxiliary conjectures, it is nevertheless true that the famous Quine-Duhem problem presents itself, in that if convergence fails, all we can say with confidence is that *something* is wrong in the conjectured conjunction of main and auxiliary statements. It is well known to logicians and philosophers of science that no straightforward automatic touchstone procedure exists for identifying the culprit in such circumstances, although very few working scientists would accept the strong form of the Quine-Duhem thesis sometimes stated, that *since* the formal logic of a negated conjunction is a disjunction of negations, *therefore* a scientist's first-round choice of the likely culprit cannot be principled, "rational," or other than whimsical. Although from the standpoint of formal logic, *modus tollens* refutation of a theoretical conjunction amounts

to a disjunction of negations of the conjuncts, when we distinguish considerations of formal logic from methodology of science, we know that in ongoing scientific research some of the conjuncts are considered less problematic than others. This is true either because they flow quasi-deductively from well-corroborated background knowledge, or because the investigator has meanwhile tested them separately from the others by a suitable choice of the experimental context. These complexities have to be accepted, and they will not be troublesome to us if the four methods converge well when examined as described in the next section of the paper. The disturbing case is the one in which they do not converge, and that would lead us to look at whether some agree fairly well with each other but one of them, as for instance Dewar's Method, is markedly out of line.

But the existence of a fossil record for extant forms provides a more direct path to assessing the methodological power of my three methods. Instead of inferring that the methods have some validity because they converge on the unknown latent value of the species number, we can apply my three methods to the population for which that species number N is known, to wit, extant forms.

By applying my three methods to a suitably chosen set of orders of extant forms (qualitatively diverse and with species numbers both sizable and variable), we can ask about each method how accurately it estimates the total number of extant species of carnivores, primates, crustaceans, insectivores, and so on. We can also get an idea of how well the composite of three weighted estimates performs as estimator. There are several interesting and important questions about the three indirect methods that can be studied in this way, of which I list the more obvious and crucial ones, but I daresay others will occur to the reader:

a. What is the average accuracy of each of the three indirect methods? ("Accuracy" in this list means numerical closeness to the true species number for each order estimated.)
b. What is the variation in accuracy of each method over the different orders?
c. Are the three indirect methods more or less random in their discrepancies, or can we discern a ranking in the species numbers estimated?
d. If there is such a ranking, can one strike a reasonable average bias, so that a corrected estimate for the clearly biased method can be applied to the extinct orders, the anticipation being that this will improve the consistency of the results on that group?
e. How do some of the auxiliary conjectures fare in this situation, where they can be checked directly against the true numbers rather than indirectly, i.e., by their success when used in a derivation chain eventuating in a numerical value testable for its consistency with other numerical values? For example, the conditional probability of sandwich sizes for a specified longevity and slice-probability generates a family of graphs of expected frequencies. How well do those theoretical graphs fit the empirically computed ones for the case in which the slice-q values are accurately determinable from the observations on extant forms, and only the lower end of the longevity is problematic?

A good convergence of the numerical values of the four methods on the extinct orders would have a satisfactory Popperian corroborative effect, but it is evident that we would gain a good deal of additional confidence in the entire interlocking system of

substantive and auxiliary conjectures if we found that a methodological claim (like an instrument's precision) is well corroborated for three of the methods used on extant forms for which the true values are known, namely, that these three methods do tend to give the right answer when it is directly known to us.

Agreement of the Four Methods

We now have the estimated number of extinct Carnivora species as inferred from the Discovery Asymptote Method, the Binomial Parameters Method, the Sandwich Method, and Dewar's Method of Extant Forms. Ditto for ungulates, insectivores, primates, etc. If the conjectured statistical model (reflecting the underlying historical *causal* model of the consecutive processes: fossilization + preservation + discovery) possesses moderate or good verisimilitude despite the idealizations and approximations involved, the four extinct species numbers estimated for a given order of, say, vertebrates should show a "reasonable agreement" with one another. I put the phrase "reasonable agreement" in double quotes to emphasize that this is a loose concept and that I do *not* intend to advocate performing a traditional statistical significance test. That would be inappropriate since it asks a question to which we already have an answer, namely, are the methods precisely equivalent in the unknown quantities they are estimating and have causally arisen from? The answer to this question is surely negative. The notion of "reasonable agreement" is not an exact notion even when stated in terms of tolerances, but biological and earth scientists are accustomed (more than social scientists) to dealing with that notion. It is scientifically a more useful and powerful notion by far than the mere trivial refutation of the null hypothesis that two population values are exactly equal—which, of course, in the life sciences they never are (see Meehl 1967; Lykken 1968; Morrison & Henkel 1970; Meehl 1978).

It may be objected by purists that, lacking a precise measure of agreement among the four numerical values (such as provided by the traditional statistical significance testing approach), one cannot say anything rational about the data. Philosophical replies to such perfectionism aside, one can only appeal to the history of the exact sciences, where for several centuries before the invention of statistical significance testing, or even "exact" procedures for setting up confidence intervals around point estimates (such as the physicists' and astronomers' probable error), great advances were made in these disciplines by studying graphs and noting reasonable numerical concordances in tables. Further, in evaluating a procedure for estimating unknown quantities one must evaluate the knowledge provided by the estimates, however crude, in *relation to our state of knowledge without them*. At the present time, almost nothing numerical has been said (given the neglect of Dewar's work) about the completeness of the fossil record. The standard comment is that it is surely very incomplete, a statement necessitated by the absence of the many thousands of transitional forms that neo-Darwinian theory requires to have existed but that are not found in the presently available fossil record. That being our current epistemic situation, even a ballpark estimate of so scientifically important a number as the "completeness index" is an advance over no quantitative estimate at all.

Nothing prevents us from treating the completeness index as a proportion or percentage and attaching an estimated standard error to it in the traditional way if we believe that is helpful. The reader may think it worthwhile to perform a significance test between completeness indices estimated by different methods. Because of my distaste for the

widespread abuse of such procedures in the behavioral sciences, I would not myself prefer to do that (see Meehl 1978); but no doubt some others would. If it makes anybody feel better to calculate a significance test over the completeness indices obtained by the four methods, he may do that. At this stage, in this place, I wish primarily to emphasize that if we could set only the most extreme upper and lower bounds upon estimated completeness, we would have made a material advance on the present state of knowledge.

For example, the consensus among evolutionists would be that the fossil record is very considerably less than half complete. Now suppose that the completeness indices obtainable by the four methods for, say, order Carnivora were numbers like 10 percent, 12 percent, 20 percent, 31 percent. These four percentages are discouragingly dispersed from the standpoint of a statistician emphasizing a precise point value; but they are important nevertheless, because they tell us that four nonredundant methods of estimating the completeness of the record for carnivores are well under one-half. Imagine instead that we found four completeness indices that were, though varying considerably among themselves, all higher than 50 percent. (Thus, for instance, the completeness index for land mammals in Dewar's Method was found by him to be 60 percent—a proportion very much higher than anyone would expect, and one that obviously has grave implications for the received doctrine of evolutionary slowness, i.e., the number of transitional forms, and their species longevities, with which the evolutionary process occurs.) This would be an important finding despite the variance over four methods. In a nutshell what I am saying is something that scientists working in such semi-speculative fields as cosmology have long been familiar with accepting as a part of their intellectual burden of uncertainty for studying a particular subject matter in which precision is difficult, to wit, sometimes it's worth estimating a quantity even within a few orders of magnitude, if the quantity is important enough, and if the orders of magnitude, from another point of view, differ only slightly in relation to the possible range we would assign in a state of complete Bayesian ignorance. A completeness index of .50 would be an order of magnitude larger than what paleontologists seem generally to consider "reasonable," presumably because of the sparsity of quasi-complete fossil phylogenies. Everyone interested in the theory of evolution would surely like to know whether the following statement can be considered more or less corroborated by reasonable estimates independent of each other: "The *majority* of species of animals with hard parts that have ever lived on the earth are known to us from the fossil record, at the present time." This would surely be a surprising generalization if corroborated by the proposed four methods. Given a fairly consistent set of completeness indices that were uniformly over 50 percent for different kinds of animals with hard parts —even if the percentages for a given order varied considerably, as between, say, 55 percent and 90 percent—we would possess a numerical result whose importance for evolutionary theory can hardly be exaggerated.

There is, however, a simple, straightforward statistical approach to evaluating the agreement, which I set forth briefly because it yields a rough quantitative measure of goodness of agreement. Most of us would find even a crude measure more reassuring than the mere verbal characterization "reasonably good agreement" among sets of four numbers. Let the reader visualize the state in which our data are now summarized, with four columns (corresponding to the four methods) of numbers, each row corresponding to an order of extinct animals. The numbers appearing in this matrix are not completeness percentages but estimated total extinct species frequencies N_a, N_b N_c, ..., based upon the

completeness estimations for *a* = Carnivora, *b* = Primates, *c* = Insectivora, etc. We do it this way because the variation in number of species for different animal kinds is what we are studying when we ask whether the four methods agree to a closeness better than what one could plausibly attribute to "mere coincidence." The underlying reference base one tacitly presupposes in referring, however loosely, to "degree of agreement" among different conjectural ways of estimating one and the same underlying numerical value in nature, is always some prior information—however vague and common-sensical—about the antecedently expected range of numerical values for the kind of quantity being studied, as measured in the units that have been adopted. We need not be strict Bayesians in our theory of inductive inference (certainly not subscribers to the strong form of subjective/personalist probability theory) to insist upon the importance of background knowledge. We unavoidably rely on background knowledge in evaluating whether a convergence of two or more estimators of a theoretically conceived numerical quantity agree well enough that their agreement cannot plausibly be regarded as some kind of a happenstance or lucky accident. (From the Popperian standpoint, the accident is unlucky, since it misleads us by supplying a strong corroborator of a causally unrelated theory.)

I should emphasize that I do not here raise some novel philosophy of science of my own. I simply call attention to a fact that is common throughout the history of all quantitative empirical disciplines, to wit, the scientist develops conviction about the accuracy of his methods in the context of whatever theory (or even theory sketch) he is using to articulate and defend a method, by noting that numbers agree well. If one does not rely upon the traditional test of significance—it is inappropriate when one already knows that the theory is an idealization, that the instruments are fallible and may even contain slight systematic biases, along with random errors of measurement and sampling—then how "close" two or more numerical estimates are judged to be must obviously depend upon some antecedent knowledge, or at least expectation, as to their "ordinary range." The obvious example is the case of measuring a simple length, where whether a discrepancy of so many units in some physical measurement scale is counted for or against the accuracy of the measuring procedure in the context of the conjectured theory will depend entirely upon the range of "empirically plausible" lengths as expressed in those units. Example: If I tell you that my theory of genetic control of growth in the elephant predicts that the trunks of baby elephants born of the same mother tend to be within six inches of each other in length, you will be unimpressed. If I tell you that they are usually within two millimeters of each other, you either will be impressed or think I am faking the data. The only basis for this difference in attitude is your prior knowledge about how big, roughly, a baby elephant's trunk is and how organisms tend to show variation. In astrophysics and cosmology, it may be worthwhile to estimate a quantity within several orders of magnitude. In other branches of physics, it may be pointless to do an experiment whose accuracy of measurement is not at least 99.99 percent, and so on.

In the light of these considerations, a reasonable way to deal with our four columns of estimated species numbers for different orders of extinct animals is to express them not as percentages or proportions (the completeness index) but as estimated absolute frequencies of species per order. Thus we are taking advantage of the presumed wide range of differences in the number of species subsumed under a particular order to get a handle on the "more-than-coincidence agreement" question. Keep in mind that we are contrasting a claim of reasonable agreement with the counter-conjecture that the four methods are

hardly any good at all, that they are nearly worthless for the purpose of estimating the mysterious number "C = completeness of the fossil record." If that pessimistic view is substantially correct, a set of calculations concerning the true unknown number of extinct species N_a for order a might as well be drawn from a hat, *and the hat to draw it from is the whole range of these species numbers*, *as generated by these invalid estimation procedures*. One way of looking at it is that in order to know whether estimates are in decent agreement on the species numbers for carnivores, insectivores, and rodents, we would like to know the basic range of species numbers for taxa at this level of classification. Thus an average disagreement of 20 species over the four methods would be discouraging if the average estimated number of species in an order ran around 40 and varied from that number down to 0; whereas if the number of species representing a particular order were characteristically tallied in the hundreds, an average disagreement of 20 species would be encouraging, given the present weak state of knowledge.

One simple and intuitively understandable expression of the degree of agreement among four measures "of the same thing" comes from classical psychometrics. It was invented three quarters of a century ago by the British engineer and psychologist Charles Spearman, who developed the so-called tetrad difference criterion for testing the psychological hypothesis that the pairwise correlations among mental tests could be attributed to the causal influence of one underlying common factor, Spearman's g. It was subsequently shown by L. L. Thurstone in the matrix development of multiple factor analysis (a generalization of Spearman's idea to the case of more than one latent causal variable determining the pattern of mental test scores) that the tetrad difference criterion is the limiting case of evaluating the rank of a matrix. The arrangement of correlation coefficients in Spearman's tetrad difference criterion is seen, from the standpoint of matrix algebra, to be the expansion of a second-order determinant in the larger matrix of correlations among many tests. Hence the vanishing of the tetrad differences tells us that the second-order determinants vanish (the corresponding second order matrices being of unit rank); and that is what Spearman's method requires if one factor suffices to account for all the relationships.

Treating each species number in a column as a nonredundant estimator of the true unknown species number, the correlations between columns in our table summarize pairwise agreement of the methods. Given four methods, there are six such inter-method correlation coefficients. As they are all estimates of the same true quantity [N_t = Number of extinct species in order t] that varies over different orders, then even if these estimates are highly fallible, the six correlations are attributable to the influence of the one latent common factor, to wit, the true unknown extinct species number N_t, so the tetrad difference equations should be satisfied. They read:

$$\begin{aligned} r_{12}\,r_{34} - r_{13}\,r_{24} &= 0 \\ r_{12}\,r_{34} - r_{14}\,r_{23} &= 0 \\ r_{13}\,r_{24} - r_{14}\,r_{23} &= 0 \end{aligned} \qquad [39]$$

There are statistical significance tests for the departure of these equations from the idealized value 0, but one does not expect them to be exactly satisfied (any more than they were *exactly* satisfied in Spearman's psychometric model). One would like to know whether they are reasonably close to 0. If they are nearly satisfied, we have reason to think that our conjecture that each method is an informative way of estimating the

unknown number of species has verisimilitude. Of course we would also expect the pairwise correlations to be high, because these correlations represent the extent to which the four methods agree. The average amount of their disagreement is thus being compared (via the inner structure of a Pearson *r*) to the variation of each method over the different orders of animals. If the tetrad difference equations are almost satisfied, it is then possible to assign a weight or loading to each method. One method may, so to speak, be "doing the best job of whatever all four of them are doing," as represented by the pattern of the six correlations of the methods, taken pairwise. One can then combine these weights to reach an improved estimate of the species number for a given order by multiplying the standardized number for each order by its common factor weight and adding these four products. What that amounts to is relying more on one measure than another, while taking them all into account, each one being weighted as befits its apparent validity for the latent quantity. The loading or saturation of a given indicator among the four is obtained by one of several approximate (and nearly equivalent) formulas, the simplest of which is

$$[40] \qquad a_{kg} = \frac{\sum r_{ks}}{\sqrt{\sum r_{ij}}}$$

where one divides the sum of indicator *k*'s correlations with the other three indicators by the square root of the sum of all correlations in the table (Thurstone 1947, pp. 153, 279; Gulliksen 1950, pp. 343-345; Harman 1960, pp. 337-360).

I do not urge that this is the optimal way to analyze these data, but with four measures the tetrad method springs naturally to mind. Other procedures would be defensible, such as the so-called generalized reliability coefficient developed many years ago by the psychologist Paul Horst (1936). I suppose most contemporary statisticians not oriented to psychometrics would prefer some generalized breakdown of components in the analysis of variance where the total sum of squares of species numbers is assigned to methods, orders, and methods × orders interaction (see, e.g., Hoyt 1941). If the methods proposed have even a crudely measured degree of reliability and construct validity as shown by their agreement, there will then be plenty of time to select the optimal method of expressing disagreement, and of combining the statistics into one "best available estimate" of the species number for each order. In the present state of the art, most of us would be happy to contemplate even a table of simple "percent deviations" of the methods from each other and their unweighted means.

Summary

Given a neo-Popperian philosophy of science, when one estimates values of unobserved theoretical variables (latent, underlying, causative, or historical) in reliance on certain mediating assumptions, these assumptions are treated as *auxiliary conjectures*, more or less problematic. Being problematic, they should not be left in the status of "assumptions," in the strong sense of required postulates that we hope are true but have complete liberty to deny. It is argued that such idealizing auxiliary conjectures can motivate four distinct, nonredundant, observationally and mathematically independent methods for estimating the degree of completeness of the fossil record as known to paleontologists at a given time. We can ascertain the closeness of numerical agreement ach-

ieved by these four nonredundant methods when each is used to estimate the total number of species, known and unknown, of a given taxon at intermediate level in the taxonomic tree (e.g., a family, order, or class). This procedure subjects the auxiliary conjectures rationalizing each method to a relatively severe Popperian test, because if the methods are poor estimators, having little or no relation to the true unknown species number for each order, there is no reason why they should have any appreciable tendency to converge. On this kind of reasoning, it is urged that paleontologists need not—I would say, *should* not—content themselves with stating that the fossil record is "very incomplete." This posture has allowed philosophic critics, including Sir Karl Popper, to argue that the theory of organic evolution is not, as it is claimed to be by geologists and systematic zoologists, a scientific theory at all, but merely a fruitful metaphysical speculation. Four methods are described, three of which—the Discovery Asymptote Method, the Binomial Parameters Method, and the Sandwich Method—were devised by the present author, and the fourth had been invented by Douglas Dewar. It is suggested that the four methods could be applied to some classes of data already catalogued in paleontology, and that future research ought to be conducted in such a way as to make available the raw data necessary for applying the methods to the fossil record for animals of many different kinds. If the methods show reasonable agreement (as measured by any of several statistical procedures), each method can be assigned a weight in proportion to its validity as an indicator of the latent quantity: "Total extinct species of the order." A composite estimate obtained by summing these weighted estimates from the four methods can then be used to calculate a completeness index for each order studied. In addition to the validity argument from numerical convergence, the trustworthiness of the author's three indirect methods can also be checked more directly by applying them to extant forms, where the true species numbers are known. It is urged that crude approximate results from these methods would represent a material advance on the present state of paleontology, where only very broad and untested qualitative statements concerning the record's gross inadequacy are presently available. It is also argued that approximate measures of consistency are of much greater scientific value than traditional use of statistical significance tests, since we know that the auxiliary conjectures are idealizations, false if taken literally, so that showing this is pointless.

Appendix I

It would be surprising, given what we surely must presume about minuscule find-probabilities, to discover that the fossil specimen frequencies per species are distributed too symmetrically ($p \simeq q$) for the Poisson to provide a good empirical fit. Such a result, locating the modal fossil count clearly to the left of our empty column at $k = 0$, would in itself strongly discorroborate the received conjecture that the record is grossly incomplete. But if the Poisson did appear to give a rather poor fit, a somewhat more complicated Binomial Parameters procedure would be needed. I present it here briefly in two variants, without any conviction as to its optimality, and expecting it never to be needed.

The problem is to estimate the latent parameters of a binomial that characterizes the find-probability in relation to digs. The latent function conjectured to generate our observed distribution of fossil-specimens-per-species is simply $N(p + q)^n$ where p = Find-probability, n = Number of digs, N = Total number of extinct Carnivora species, found and unfound

The unknown N cancels when we form ratios of successive terms. t_1/t_2, t_2/t_3, t_3/t_4, …, of the expansion of this binomial. If we then take ratios of these term-ratios, the latent parameters (p, q) also cancel, and the ratio-ratios (from the low-find end) run

$$[41] \qquad \frac{1}{2}\frac{n-1}{n}, \frac{2}{3}\frac{n-2}{n-1}, \frac{3}{4}\frac{n-3}{n-2}, \ldots$$

Which for large n (= number of digs, not empirically known but surely large enough) involves n-ratios of form $(n - r)/(n - r + 1) \simeq 1$. Hence the ratios of term ratios are closely approximated by simple fractions involving small integers (at the low find frequency end) and the last one, involving the missing term at $k = 0$ (no fossil found) is

$$[42] \qquad R = (t_1 / t_2) / (t_2 / t_3) = \frac{1}{2}\left(\frac{n-1}{n}\right)$$

very near to ½, surely an error of less than 1 percent. So we solve in [41] for the missing number of species at $k = 0$ finds.

Another possibility utilizes more distribution of information but I fear could be quite unstable, so the added information might not pay off. The ratios of successive terms, although the equations for them will be empirically inconsistent, could be solved for the parameters p, n. Then the missing term at $k = 0$ is calculated from these. Taking this value as a first approximation to the missing item, we compute the empirical skewness and kurtosis of our find-distribution. These numerical values characterizing the empirical distribution shape are then employed with the standard formulas for skewness and kurtosis of a binomial, two equations soluble for the two parameters (p, n) (Cramér 1946, p. 195; Kenney 1939, p. 15). This gives a second approximation to them, and permits a revised estimate of the $k = 0$ probability. We iterate until the values settle down.

Appendix II

Given the expression Equation [32] for the probability of a specified size sandwich arising from longevity L, we can get the probability P_{sL} of finding a sandwich (any size) from antecedent condition L. This is the sum over all sandwich thicknesses k of the terms $p(k/L)$ as just obtained, i.e.,

$$[43] \qquad P_{sL} = \sum_{k=1}^{L-2} (L-k-1) p^2 q^{L-k-2}$$

which is like a geometric progression with common ratio q except that each term has an integer coefficient that undergoes unit increments from term to term. Thus for $L = 6$ the sum is

$$[44] \qquad \sum = 4p^2q^3 + 3p^2q^2 + 2p^2q^1 + 1p^2q^0$$

$$[45] \qquad \sum = p^2(4q^3 + 3q^2 + 2q^1 + 1q^0$$

Call the parenthesis S', multiply by q and subtract,

$$[46] \qquad S'_4 q - S'_4 = 4q^4 - \left(q^3 + q^2 + q\right) - 1$$

and now the parenthetical sum is a 3-term G.P. (call it S_3) with its ratio q equal to its

initial term, that is, after dividing by $(q - 1)$ to get S'_4 alone on the left, we have

$$[47] \qquad S'_4 = \frac{4q^4 - S_3 - 1}{q - 1} \ .$$

The general expression for total sandwich probability is easily seen to be

$$[48] \qquad P_{sL} = p^2 S' = p^2 \frac{S_{L-3} - (L-2)q^{L-2} + 1}{1-q}$$

$$= p\left[S_{L-3} - (L-2)q^{L-2} + 1 \right]$$

We have to substitute the usual formula for a G.P. for S_{L-3},

$$[49] \qquad S_{L-3} = \frac{q\left(1 - q^{L-3}\right)}{1-q}$$

in [48], obtaining

$$[50] \qquad P_{sL} = p\left[\frac{q\left(1 - q^{L-3}\right)}{1-q} - (L-2)q^{L-2} + 1 \right]$$

which after some grouping and cancelling gives us

$$[51] \qquad P_{sL} = (L-2)q^{L-1} - (L-1)q^{L-2} + 1$$

This is a fairly simple expression for sandwich-probability (any size) as a function of the slice-*im*probability q_i for a given longevity L_i.

We note that this result can be obtained more directly (and easier intuitively) by subtraction, beginning with the probability of *not* finding a sandwich (of any size) from condition L. There are three non-sandwich cases:

	Probability
Not found at all	q^L
One bread slice	$L\,p\,q^{L-1}$
Two adjacent bread slices (no "inner" tally available)	$(L-1)\,p^2\,q^{L-2}$

The sandwich probability is the complement of the sum of these three nonsandwich probabilities,

$$[52] \qquad P_{sL} = 1 - \left(q^L + Lpq^{L-1} + \left(L-1\right)p^2 q^{L-2} \right)$$

which with a little manipulation readily yields the expression [51] reached positively.

These longevity-conditional probabilities cannot be quickly evaluated from our observed within-sandwich slice-probabilities because the empirical slice-p's do not correspond to the latent slice-p variable occurring in the conditional probability formula.

References

Cramér, H. (1946) *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Cronbach, L.J. & Meehl, P.E. (1953) Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Darwin, C. (1959) On *the origin of species by means of natural selection*. London: John Murray.

Dewar, D. & Shelton, H.S. (1947) *Is evolution proved?* London: Hollis & Carter

Dobzhansky, T. (1940) Catastrophism versus evolutionism. *Science*, *92*, 156-155.

Eldredge, N. & Gould, S.J. (1972) Punctuated equilibria: An alternative to phyletic gradualism. In T.J.M. Schopf (Ed.), *Models in paleobiology* (pp. 82-115). San Francisco: Freeman, Cooper & Company.

Feller, W. (1957) *An introduction to probability theory and its applications* (2nd ed.). New York: Wiley.

Golden, R.R. & Meehl, P.E. (1978) Testing a single dominant gene theory without an accepted criterion variable. *Annals of Human Genetics*, *41*, 507-514.

Golden, R.R. & Meehl, P.E. (1979) Detection of the schizoid taxon with MMPI indicators. *Journal of Abnormal Psychology*, *88*, 217-233.

Golden, R.R. & Meehl, P.E. (1980) Detection of biological sex: An empirical test of cluster methods. *Multivariate Behavioral Research*, *15*, 475-496.

Goldschmidt, R.B.G. (1940) *The material basis of evolution*. New Haven, CT: Yale University Press.

Gulliksen, H. 1950. *Theory of mental tests.* New York: John Wiley & Sons.

Harman, H.H. (1960) *Modern factor analysis.* Chicago: University of Chicago Press.

Hays, W.L. (1973) *Statistics for the social sciences* (2nd ed.). New York: Holt. Rinehart & Winston.

Horst, P. (1936) Obtaining a composite measure from different measures of the same attribute. *Psychometrika*, *1*, 53-60.

Hoyt, C. (1941) Test reliability obtained by analysis of variance. *Psychometrika*, *6*, 151-160.

Hull, C.L. (1943) *Principles of behavior*. New York: Appleton-Century.

Kenney, J.F. (1939) *Mathematics of statistics*. New York: D. Van Nostrand Company.

Kyburg, H.G.E. (1969) *Probability theory.* Englewood-Cliffs, NJ: Prentice Hall.

Lewis, D. (1960) *Quantitative methods in psychology*. New York: McGraw-Hill.

Lykken, D.T. (1968) Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-159.

Meehl, P.E. 1965. *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report No. PR-65-2. Minneapolis.

Meehl, P.E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.

Meehl, P.E. (1968) *Detecting latent clinical taxa. II: A simplified procedure*, *some additional hitmax cut locators*, *a single-indicator method*, *and miscellaneous theorems*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report No. PR-68-4. Minneapolis.

Meehl, P.E. (1970) Nuisance variables and the ex post facto design. In M. Radner and S. Winokur (Eds.), *Minnesota Studies in the Philosophy of Science*, *Vol. IV: Analyses of theories and methods of physics and psychology* (pp. 373-402). Minneapolis: University of Minnesota,

Meehl, P E. (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834

Meehl, P.E. (1979) A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment*, *43*, 563-581.

Meehl, P.E. & Golden, R.R (1982) Taxometric methods. In P. Kendall and J.N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127-181). New York: Wiley.

Morrison, D.E. & Henkel, R.E. (Eds.) (1970) *The significance test controversy.* Chicago: Aldine.

Perrin, J. (1910) *Atoms*. New York: D. Van Nostraud Company.

Popper, K.R. (1962) *Conjectures and refutations*. New York: Basic Books.

Popper, K.R. (1972) *Objective knowledge: An evolutionary approach.* Oxford: Oxford University Press.

Popper, K.R. (1974) Intellectual autobiography. In P.A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 3-181). LaSalle, IL: Open Court.

Segré, E. (1980) *From X-rays to quarks: Modern physicists and their discoveries*. San Francisco: W.H. Freeman & Company.

Thurstone, L.L. (1947) *Multiple-factor analysis*. Chicago: University of Chicago Press.

Uspensky, J.V. (1937) *Introduction to mathematical probability*. New York: McGraw-Hill.

Yule, G.U. & Kendall, M.G. (1940) An *introduction to the theory of statistics.* London: Charles Griffin.

pdf by LJY, November 2019