Logic, Methodology and Philosophy of Sciences III,
van Rootselaar and Staal. ed., © *North-Holland Publ. Cy., Amsterdam - 1968*

#325

ON JUDGING THE PLAUSIBILITY OF THEORIES

H.A. SIMON*

Carnegie-Mellon University, Pittsburgh, Pennsylvania, USA

1. It is a fact that if you arrange the cities (or, alternatively, the metropolitan districts) of the United States in the order of their population in 1940, the population of each city will be inversely proportional to its rank in the list (see fig. 1). The same fact is true of these cities at other census dates – back to the first census – and for the cities of a number of other countries of the world.

It is a fact that if you arrange the words that occur in James Joyce's Ulysses in the order of their frequency of occurrence in that book, the frequency of each word will be inversely proportional to its rank in the list (see fig. 2). The same fact is true of the other books in English whose word frequencies have been counted (except, possibly, *Finnegan's wake*), and it is true of books in most other languages (although not books in Chinese).

What do I mean when I say these are "facts"? In a way, it seems incorrect to speak in this way, since none of my "facts" is literally and exactly true. For example, since there were 2034 cities over 5000 population in the United States in 1940, the alleged "fact" would assert that there were therefore one half as many, 1017, over 10000 population. Actually, there were 1072. It would assert that there were one tenth as many, 203, over 50000 population; actually, there were 198. It would assert that the largest city, New York, had a population just over ten million people; actually, its population was seven and one half million. The other "facts" asserted

---

* This work was supported in part by Public Health Service Research Grant MH-07722 from the National Institutes of Mental Health.

I should like to dedicate this essay to the memory of Norwood Russell Hanson, in acknowledgment of my debt to his *Patterns of discovery*. His work did much to reestablish the notion that the philosophy of science must be as fully concerned with origins of scientific theories as with their testing – indeed that the two are inextricably interwoven. His reconstruction of Kepler's retroduction of the laws of planetary motion will long serve as a model of inquiry into the history and philosophy of science.

above, for cities and words, hold only to comparable degrees of approximation.

At the very least, one would think, the statements of fact should be amended to read "nearly inversely proportional" or "approximately inversely proportional" rather than simply "inversely proportional". But how near is "nearly", and how approximate is "approximately"? What degree of deviation from the bald generalization permits us to speak of an approxi-

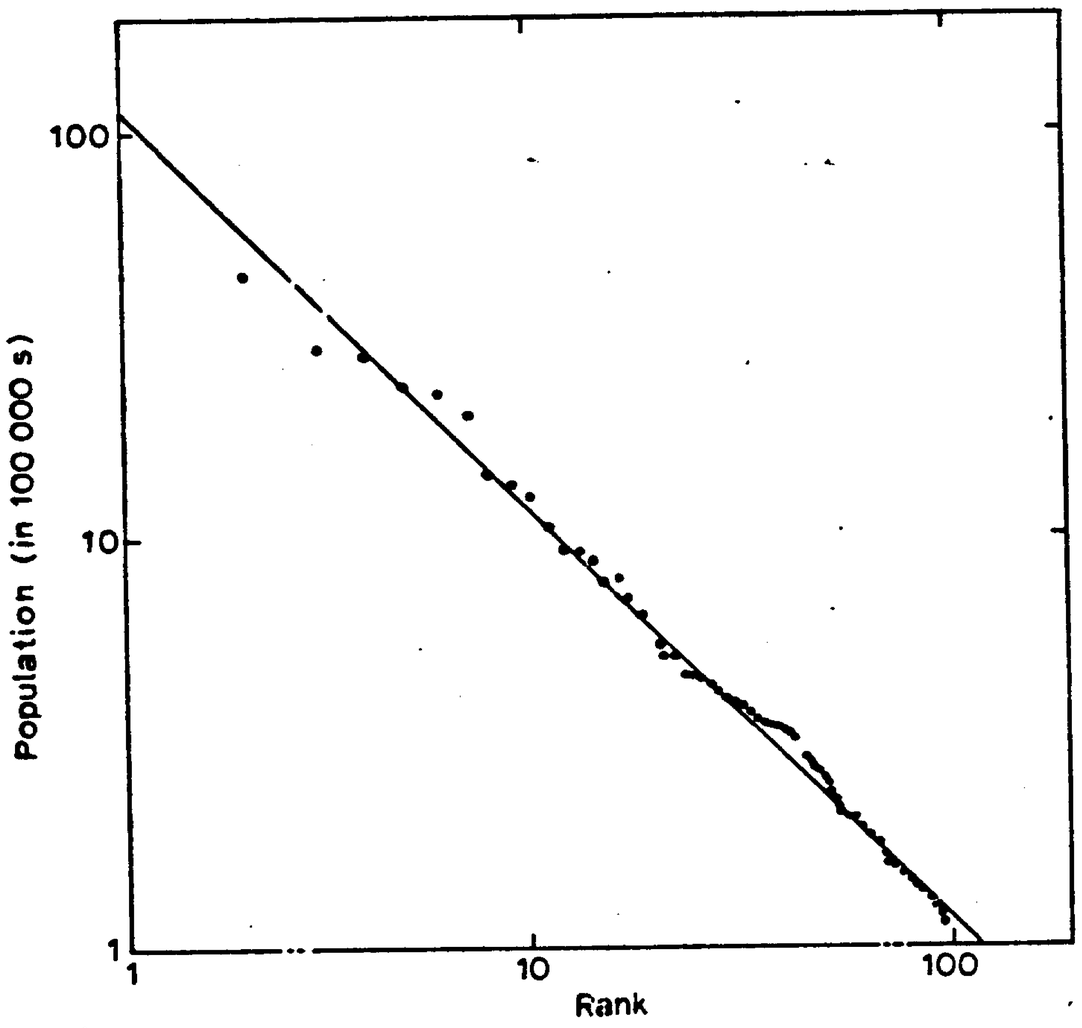

Fig. 1. Hundred largest U.S. cities, 1940 (ranked in order of decreasing size).

mation to the generalization rather than its disconfirmation? And why do we prefer the simple but approximate rule to the particular but exact facts?

2. It is well known – at least among mathematical statisticians – that the theory of statistical tests gives us no real help in choosing between an approximate generalization and an invalid one.[1] By imbedding our

[1] For a brief, but adequate statement of the reasons why "literally to test such hypotheses ... is preposterous", see SAVAGE [1954] pp. 254–256. Since such tests are still reported frequently in the literature, it is perhaps worth quoting SAVAGE [1954] p. 254 at slightly greater length: "The unacceptability of extreme null hypotheses is perfectly well known; it is closely related to the oftenheard maxim that science disproves, but never proves,

generalization in a probability model, we can ask: If this model describes the real "facts" what is the probability that data would have occurred at least as deviant from the generalization as those actually observed? If this probability is very low – below the magic one per cent level, say – we are still left with two alternatives: the generalization has been disconfirmed, and is invalid; or the generalization represents only a first approximation to the true, or "exact" state of affairs.

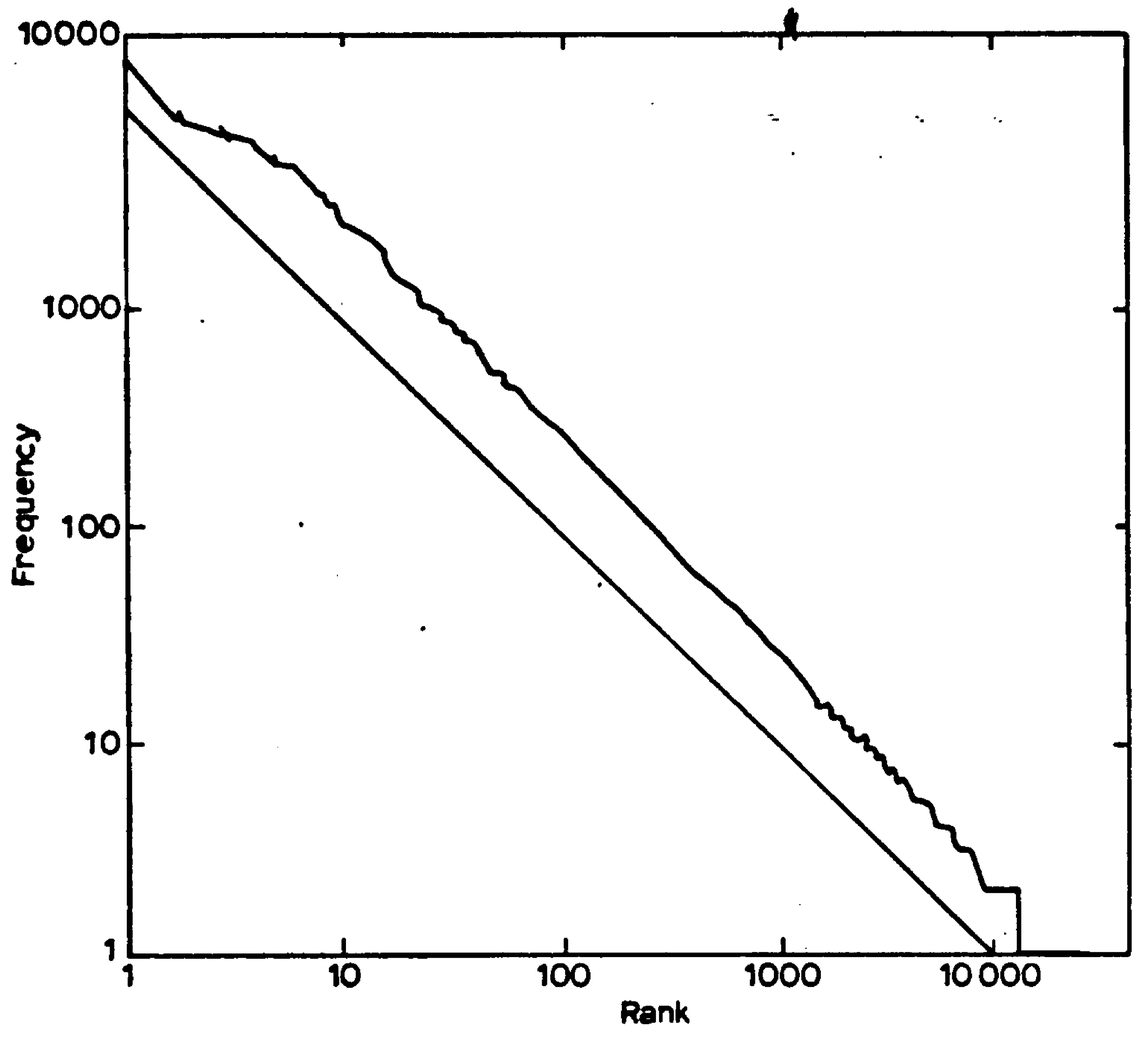

Fig. 2. Words occurring in Joyce's Ulysses (ranked by frequency of occurrence).

Now such approximations abound in physics. Given adequate apparatus, any student in the college laboratory can "disconfirm" Boyle's Law – i.e., can show that the deviations of the actual data from the generalization that the product of pressure by volume is a constant are too great to be dismissed as "chance". He can "disconfirm" Galileo's Law of Falling Bodies even

---

hypotheses. The rôle of extreme hypotheses in science and other statistical activities seems to be important but obscure. In particular, though I, like everyone who practices statistics, have often 'tested' extreme hypotheses, I cannot give a very satisfactory analysis of the process, nor say clearly how it is related to testing as defined in this chapter and other theoretical discussions".

more dramatically – the most obvious way being to use a feather as the falling body.

When a physicist finds that the "facts" summarized by a simple, powerful generalization do not fit the data exactly, his first reaction is *not* to throw away the generalization, or even to complicate it by incorporating additional terms. When the data depart from $s=\frac{1}{2}gt^2$, the physicist is not usually tempted to add a cubic term to the equation. (It took Kepler almost ten years to retreat from the "simplicity" of a circle to the "complexity" of an ellipse.) Instead, his explorations tend to move in two directions: (1) toward investigations of his measurement procedures as possible sources of the discrepancies; and (2) toward the identification of other variables associated with the deviations. These two directions of inquiry may, of course, be interrelated.

In his concern with other variables, the physicist is not merely or mainly concerned with "control" in the usual sense of the term. No amount of control of air pressure, holding it, say, exactly at one atmosphere, will cause a feather to obey Galileo's Law. What the physicist must learn through his explorations is that as he decreases the air pressure on the falling body, the deviations from the law decrease in magnitude, and that if he can produce a sufficiently good vacuum, even a feather can be made to obey the law to a tolerable approximation.

In the process of producing conditions under which deviations from a generalization are small, the scope of the generalization is narrowed. Now it is only claimed to describe the facts "for an ideal gas", or "in a perfect vacuum". At best, it is asserted that the deviations will go to zero in the limit as the deviation of the actual experimental conditions from the "ideal" or "perfect" conditions goes to zero.

At the same time that the breadth of the empirical generalization is narrowed by stating the conditions, or limiting conditions, under which it is supposed to hold, its vulnerability to falsification is reduced correspondingly. Since this is a familiar feature of theorizing in science, I will not elaborate on the point here.

Occasionally, an empirical generalization is abandoned, after innumerable attempts to tidy it up have failed. Bode's Law, that the successive distances of the planets from the Sun constitute an approximate geometric series, is an example of a regularity now regarded as perhaps "accidental", through failure to discover limiting conditions that would regularize it, or underlying processes that would account for it. Newton's Laws are *not* an example, for they were saved (a) by limiting them to conditions where velocities are

low relative to the velocity of light, and (b) by showing that just under those conditions they can be derived in the limit from the more general laws of Relativity.

From these, and many other examples, we can see what importance the physical and biological sciences attach to finding simple generalizations that will describe data approximately under some set of limiting conditions. Mendel's treatment of his sweet-pea data, as reflecting simple ratios of 3 to 1 in the second-generation hybrids, is another celebrated illustration; as is Prout's hypothesis (uneasily rejected by chemists for several generations until its exceptions were explained by the discovery of isotopes) that all atomic weights are integral multiples of the weight of the hydrogen atom. All of these examples give evidence of strong beliefs that when nature behaves in some unique fashion – deals a hand of thirteen spades, so to speak – this uniqueness, even if approximate, cannot be accidental, but must reveal underlying lawfulness.

3. Let us return to city sizes and word frequencies. We have described the law-finding process in two stages:

(1) finding simple generalizations that describe the facts to some degree of approximation;

(2) finding limiting conditions under which the deviations of facts from generalization might be expected to decrease.

The process of inference from the facts (the process called "retroduction" by Peirce and Hanson[2]) does not usually stop with this second stage, but continues to a third:

(3) explaining why the generalization "should" fit the facts. (Examples are the statistical-mechanical explanation for Boyle's Law or Boyle's own "spring of the air" explanation, and Newton's gravitational explanation for Galileo's Law.)

Before we go on to this third stage, we must consider whether we have really been successful in carrying out the first two for the rank-size distributions.

Does the generalization that size varies inversely with rank really fit the facts of cities and words even approximately? We plot the data on double log paper. If the generalization fits the facts, the resulting array of points will (1) fall on a straight line, (2) with a slope of minus one.

Since we earlier rejected the standard statistical tests of hypotheses as inappropriate to this situation, we are left with only judgmental processes

---

[2] HANSON [1961] pp. 85–88.

for deciding whether the data fall on a straight line. It is not true, as is sometimes suggested, that almost *any* ranked data will fall on a straight line when graphed on doubly logarithmic paper. It is quite easy to find data that are quite curvilinear to the naked eye (see fig. 3). Since we are not committed to exact linearity but only approximate linearity, however, the conditions we are imposing on the data are quite weak, and the fact that they meet the conditions is correspondingly unimpressive. We may therefore find the evidence unconvincing that the phenomena are "really" linear in the limiting cases. The phenomena are not striking enough in this respect to rule out coincidence and chance. Should we believe the data to be patterned?

It has often been demonstrated in the psychological laboratory that men – and even pigeons – can be made to imagine patterns in stimuli which the experimenter has carefully constructed by random processes. This behavior is sometimes called "superstitious", because it finds causal connections where the experimenter knows none exist in fact. A less pejorative term for such behavior is "regularity-seeking" or "law-seeking". It can be given a quite respectable Bayesian justification. As JEFFREYS and WRINCH [1921] have shown, if one attaches a high a priori probability to the hypothesis that the world is simple (i.e., that the facts of the world, properly viewed, are susceptible to simple summarization and interpretation); and if one assumes also that simple configurations of data are sparsely distributed among all logically possible configurations of data, then a high posterior probability must be placed on the hypothesis that data which appear relatively linear in fact reflect approximations to conditions under which a linear law holds.

The reason that apparent linearity, by itself, does not impress us is that it does not meet the second condition assumed above – the sparsity of simple configurations. A quadratic law, or an exponential, or a logarithmic, are almost as simple as a linear one; and the data they would produce are not always distinguishable from data produced by the latter.

What is striking about the city size and vocabulary data, however, is not just the linearity, but that the slope of the ranked data, on a log scale, is very close to minus one. Why this particular value, chosen from the whole non-denumerable infinity of alternative values? We can tolerate even sizeable deviations from this exact slope without losing our confidence that it must surely be the limiting slope for the data under some "ideal" or "perfect" conditions.

We might try to discover these limiting conditions empirically, or we might seek clues to them by constructing an explanatory model for the

limiting generalization – the linear array with slope of minus one. In this way we combine stages two and three of the inference process described at the beginning of this section. Let us take this route, confining our discussion to city size distributions.

4. To "explain" an empirical regularity is to discover a set of simple mechanisms that would produce the former in any system governed by the latter. A half dozen sets of mechanisms are known today that are capable of producing the linear rank-size distribution of city populations. Since they are all variations on one or two themes, I will sketch just one of them (SIMON [1955]).

We consider a geographical area that has some urban communities as well as rural population. We assume, for the urban population, that birth rates and death rates are uncorrelated with city size. ("Rate" here always means "number per year per 1 000 population".) We assume that there is migration between cities, and net emigration from rural areas to cities (in addition to net immigration to cities from abroad, if we please). With respect to all migration, we assume: (1) that out-migration rates from cities are uncorrelated with city size; (2) that the probability that any migrant, chosen at random, will migrate to a city in a particular size class is proportional to total urban population in that class of cities. Finally, we assume that of the total growth of population in cities above some specified minimum size, a constant fraction is contributed by the appearance of new cities (i.e., cities newly grown to that size). The resulting steady-state rank-size distribution of cities will be approximately linear on a double log scale, and the slope of the array will approach closer to minus one as the fraction of urban population growth contributed by new cities approaches zero.

When we have satisfied ourselves of the "reasonableness" of the assumptions incorporated in our mechanism, and of the insensitivity of the steady-state distribution to slight deviations from the assumptions as given, then we may feel, first, that the empirical generalization can now be regarded as "fact"; and, second, that it is not merely "brute fact" but possesses a plausible explanation.

But the explanation does even more for us; for it also suggests under what conditions the linearity of the relation should hold most exactly, and under what conditions the slope should most closely approximate to one. If the model is correct, then the rank-size law should be best approximated in geographical areas (1) where urban growth occurs largely in existing cities, (2) where all cities are receiving migration from a common "pool"; and (3) where there is considerable, and relatively free, migration among all the

cities. The United States, for example, would be an appropriate area to fit the assumptions of the model; India a less suitable area (because of the relatively weak connection between its major regions); Austria after World War I a still less suitable area (because of the fragmentation of the previous Austro-Hungarian Empire, see fig. 3). I do not wish to discuss the data here beyond observing that these inferences from the model seem generally to be borne out.

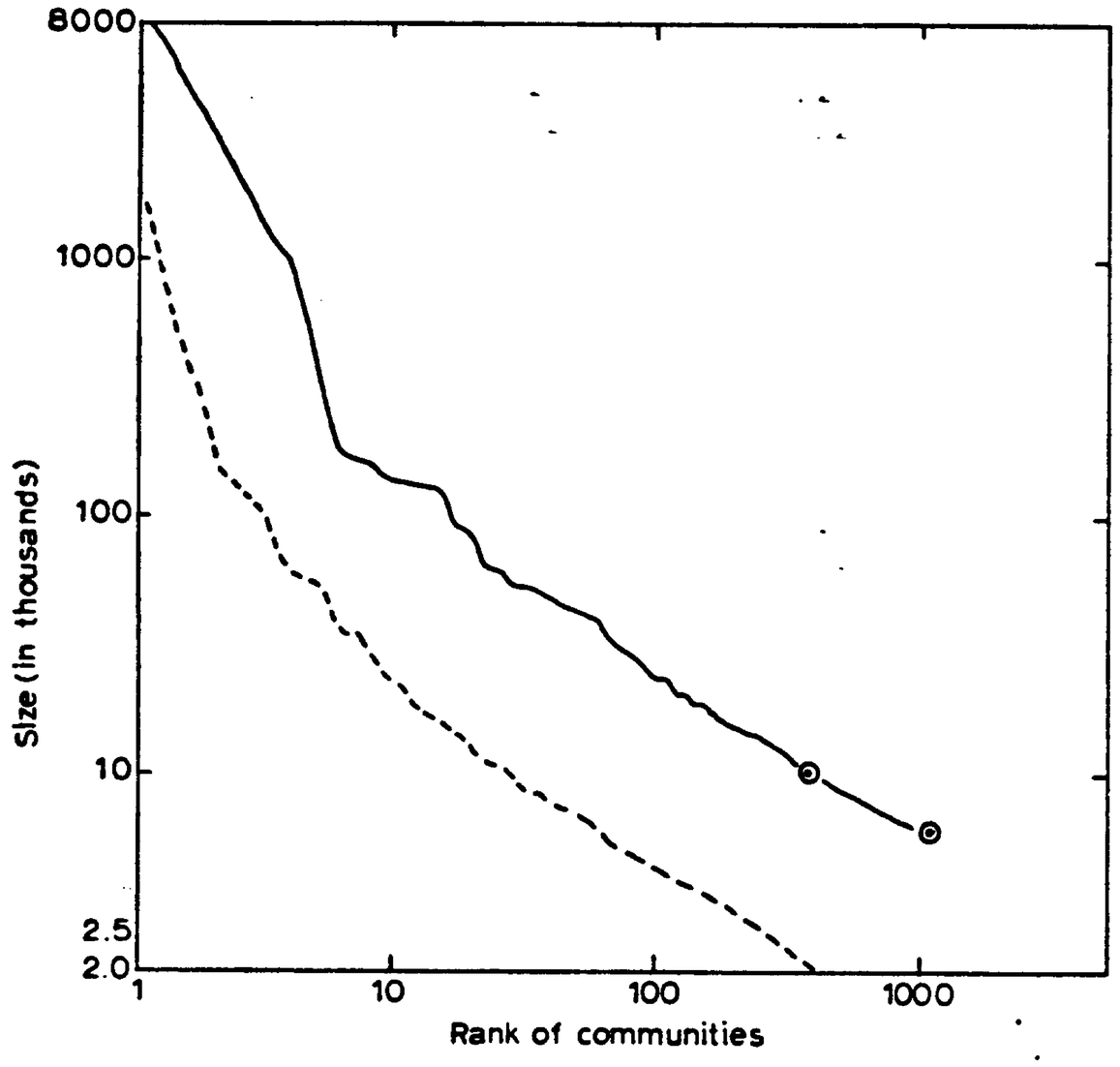

Fig. 3. Rank-size distribution of cities in Austro-Hungarian Empire, 1910 (——) and in Austria, 1934 (-----).

5. In our account thus far, the simplicity of the empirical generalization has played a central role. Simplicity is also an important concept in POPPER [1961][3] but Popper treats simplicity in a somewhat different way than we have done. Popper (on p. 140) equates simplicity with *degree of falsifiability*. A hypothesis is falsifiable to the degree that it selects out from the set of all possible worlds a very small subset, and asserts that the real world belongs to this subset.

[3] Especially Chapter VII.

There is a strong correlation between our intuitive notions of simplicity (e.g., that a linear relation is simpler than a polynomial of higher degree) and falsifiability. Thus, among all possible monotonic arrays, linear arrays are very rare. (They would be of measure zero, if we imposed an appropriate probability measure on the set of all monotonic arrays.) Linear arrays with slope of minus one are even rarer.

No one has provided a satisfactory general measure of the simplicity

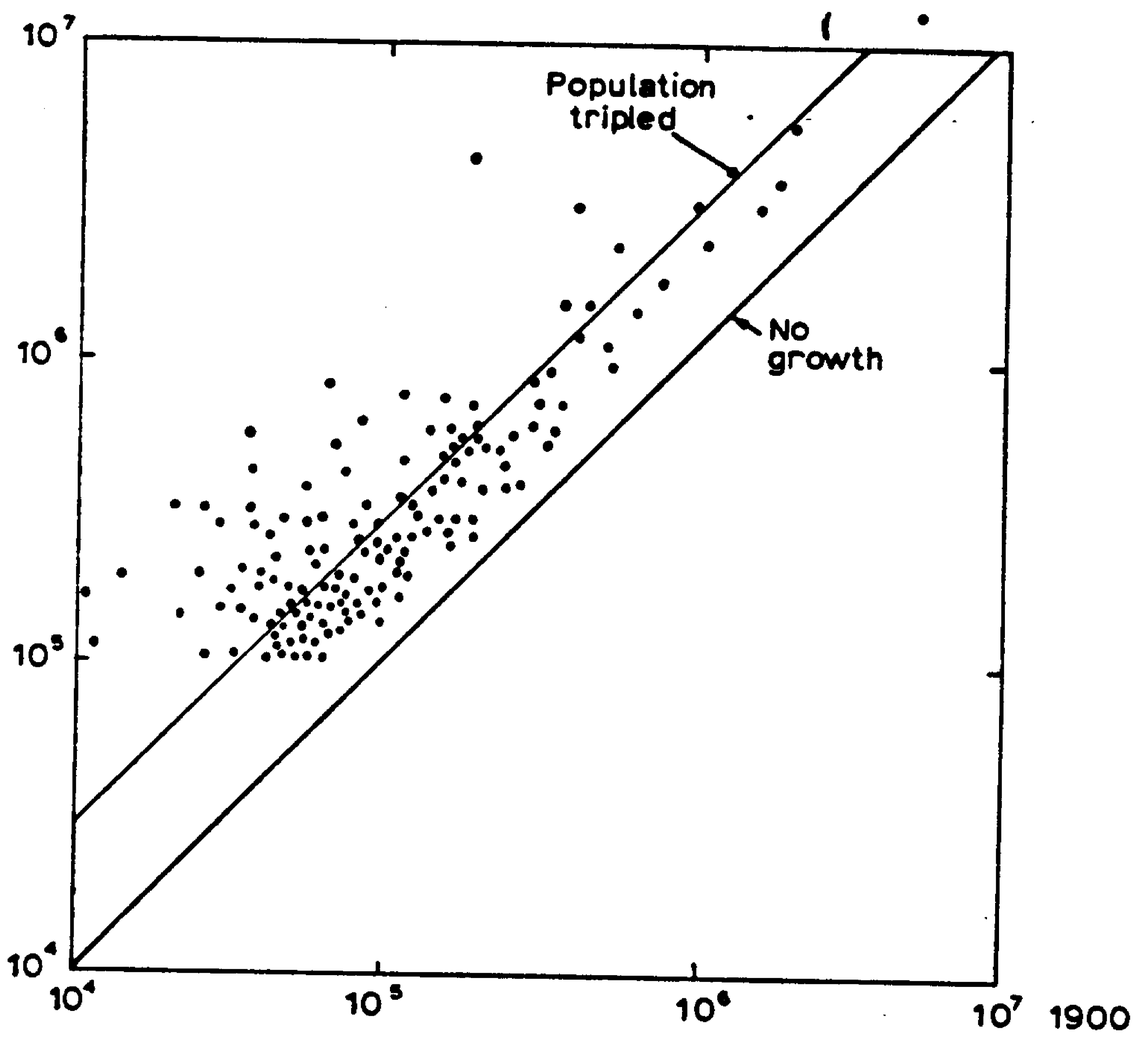

Fig. 4. Population of U.S. metropolitan districts, 1900 and 1950. (Only districts over 100000 population in 1950 are shown.)

or falsifiability of hypotheses. In simple cases, the concepts have an obvious connection with degrees of freedom: the fewer the degrees of freedom, or free parameters, the simpler and more falsifiable the hypothesis. I shall not undertake to carry the formalization of the concepts beyond this intuitively appealing basis[4].

[4] The most serious attempts at formalization are those undertaken by JEFFREYS and WRINCH [1921], and GOODMAN [1958]. I must note in passing that in his discussion of the former authors POPPER [1961] does not do justice to their technical proposal for introducing prior probabilities based on simplicity.

Notice, however, that our use of simplicity is quite different from POPPER's [1961]. Popper's argument runs like this: it is desirable that hypotheses be simple so that, if they are false, they can be disconfirmed by empirical data as readily as possible. Our argument (apparently first introduced by JEFFREYS and WRINCH [1921]) runs: a simple hypothesis that fits data to a reasonable approximation should be entertained, for it probably reveals an underlying law of nature. As Popper himself observes (POPPER [1961] p. 142, footnote*2), these two arguments take quite opposite positions with respect to the "probability" or "plausibility" of simple hypotheses. He regards such hypotheses as describing highly particular, hence improbable states of the world, and therefore as readily falsified. JEFFREYS and WRINCH [1921] (and I) regard them as successfully summarizing highly unique (but actual) states of the world, therefore as highly plausible.

Which of these views is tenable would seem to depend on which came first, the generalization or the data. If I construct generalizations, with no criterion to guide my choice except that they be simple, and subsequently apply them to data, then the simpler the generalization the more specific their description, and the less likely that they will stand up under their first empirical test. This is essentially Popper's argument.

But the argument does not apply if the generalization was constructed with the data in view. The rank-size hypothesis arises because we think to plot the data on double log paper, and when we do, it appears to be linear and to have a slope of minus one. There is no thought of using the data to falsify the generalization, for the latter has come into being only because it fits the data, at least approximately.

Now one can cite examples from the history of science of both of these alternative sequences of events. It is probably true, however, that the first sequence – generalization followed by data – seldom occurs except as a sequel to the second. The Special Theory of Relativity, for example, led to the prediction of the convertibility of mass into energy. But Special Relativity itself was based on a generalization, the Lorentz-Fitzgerald equation, that was derived to fit facts about the behavior of particles in very intense fields of force, as well as other facts about electromagnetics and the "luminiferous ether". Special Relativity did not commend itself to Einstein merely because of its "simplicity" independently of the facts to be explained (the Galilean transformations would be thought by most people to be simpler than the Lorentz).

If the generalization is just that – an approximate summary of the data – then it is certainly not falsifiable. It becomes falsifiable, or testable, when

(a) it is extended beyond the data from which it was generated, or (b) an explanatory theory is constructed, from which the generalization can be derived, and the explanatory theory has testable consequences beyond the original data.

With respect to the city size data, case (a) would arise if the rank size generalization were proposed after examining the data from the 1940 U.S. Census, and then were extrapolated to earlier and later dates, or to the cities of other countries. Case (b) would arise if we were to note that the explanatory theory of Section 4, above, has implications for patterns of migration that could be tested directly if data on points of origin and destination of migrants were available.

It should be evident that the mechanisms incorporated in the explanatory theory were not motivated by their falsifiability. They were introduced in order to provide "plausible" premises from which the generalization summarizing the observed data could be deduced. And what does "plausible" mean in this context? It means that the assumptions about birth and death rates and migration are not inconsistent with our everyday general knowledge of these matters. At the moment they are introduced, they are already known (or strongly suspected) to be not far from the truth. The state of affairs they describe is not rare or surprising (given what we actually know about the world); rather their subsequent empirical falsification would be rather surprising. What is *not* known at the moment they are introduced is whether they provide adequate premises for the derivation of the rank-size generalization.

Explaining the empirical generalization, that is, providing a set of mechanisms capable of producing it, therefore reintroduces new forms of testability to replace those that were lost by accepting the approximation to the data. Even without data on migration, the mechanism proposed to explain the city rank-size law can be subjected to new tests by constructing the transition matrix that compares the sizes of the same cities at two points of time (taking the 1900 population, say, as the abcissa, and the 1950 population as the ordinate (see fig. 4)). The explanatory mechanism implies that the means of the rows in this matrix fall on a straight line through the origin (or on a straight line of slope +1 on a log-log scale). The result (which we will expect to hold only approximately) is equivalent to the proposition that the expected growth rates are independent of initial city size.

6. In the preceding sections a model has been sketched of the scientific activities of hypothesis-generation and hypothesis-testing. The model suggests

that there are several distinct processes that lead to hypotheses being formulated, judged with respect to plausibility, and tested. One of these processes, the generation of simple extreme hypotheses from the "striking" characteristics of empirical data, fits closely the idea of JEFFREYS and WRINCH [1921] that simple hypotheses possess a high plausibility. A second process, the construction of explanations for these extreme hypotheses, takes us back to POPPER's [1961] idea that simple hypotheses entail strong and "improbable" consequences, hence are readily falsified (if false). There is no contradiction between these two views.

To elucidate further this model of the scientific process, and to reveal some additional characteristics it possesses, the remaining sections of this paper will be devoted to the analysis of a second example, this one of considerable interest to the psychology of learning and concept formation. An important question in psychology during the past decade has been whether learning is to be regarded as a sudden, all-or-none phenomenon, or whether it is gradual and incremental. One value in stating the question this way is that the all-or-none hypothesis is a simple, extreme hypothesis, hence is highly falsifiable in the sense of POPPER [1961].

The experiments of ROCK [1957] first brought the all-or-none hypothesis into intense controversy. His data strongly supported the hypothesis (even under rather strict limits on the degree of approximation allowed). Since his generalization challenged widely-accepted incrementalist theories, his experiment was soon replicated (seldom quite literally), with widely varying findings. The discussion in the literature, during the first few years after Rock's initial publication, centered on the "validity" of his data – i.e., whether he had measured the right things in his experiment, and whether he had measured them with adequate precision.

Only after several years of debate and publication of apparently contradictory findings was some degree of agreement reached on appropriate designs for testing the hypothesis. Still, some experimenters continued to find one-trial learning, others incremental learning. After several more years, the right question was asked, and the experiments already performed were reviewed to see what answer they gave[5]. The "right question", of course, was: "Under what conditions will learning have an all-or-none character?" The answer, reasonably conformable to the experimental data, commends itself to common sense. Oversimplified, the answer is that one-trial learning is likely to occur when the time per trial is relatively long, and when the

[5] POSTMAN [1963], UNDERWOOD [1964].

items to be learned (i.e., associated) are already familiar units[6]. There are the "ideal" or "perfect" conditions under which one-trial learning can be expected to occur.

7. Meanwhile, the all-or-none hypothesis was also being applied to concept attainment experiments. Important work was done in this area by Estes, by Bourne, and by Bower and Trabasso, among others. I will take as my example for discussion a well-known paper by Bower and Trabasso that Gregg and I have analysed in another context[7].

The experiments we shall consider employ an *N*-dimensional stimulus with two possible values on each dimension, and having a single relevant dimension (i.e., simple concepts). On each trial, an instance (positive or negative) is presented to the subject; he responds "positive" or "negative"; and he is reinforced by "right" or "wrong", as the case may be.

Bower and Trabasso obtain from the data of certain of their experiments an important empirical generalization: the probability that a subject will make a correct response on any trial prior to the trial on which he makes his last error is a constant. (In their data, this constant is always very close to one half, but they do not incorporate this fact in their generalization as they usually state it.) Since the generalization that the probability of making a correct response is constant is an extreme hypothesis, the standard tests of significance are irrelevant. We must judge whether the data fit the generalization "well enough". Most observers, looking at the data, would agree that they do (see fig. 5).

But Bower and Trabasso go a step further. They derive the empirical generalization from a simple stochastic model of the learning process – they explain it, in the sense in which we used that term earlier. The explanation runs thus: (1) the subject tries out various hypotheses as to what is the correct concept, and responds on individual trials according to the concept he is currently holding; (2) if his response is wrong, he tries a new concept. Two important empirical quantities are associated with

---

[6] As a matter of history, I might mention that in 1957, prior to ROCK's [1957] publication of his experiment, a theory of rote learning, designed especially to explain data that were in the literature prior to World War II (the serial position curve, the constancy of learning time per item, some of E. Gibson's experiments on stimulus similarity) had been developed by E. Feigenbaum and the author. This theory, EPAM, was sufficiently strong to predict the conditions under which one-trial learning would occur. It was not widely known among psychologists at that time, however, and had little immediate influence on the controversy. (But see GREGG, CHENZOFF and LAUGHERY [1963], also, GREGG and SIMON [1967b].)

[7] BOWER and TRABASSO [1964]; GREGG and SIMON [1967a].

the model: The probability of making a correct response prior to the last error; and the probability that any particular trial will be the trial of last error.

Now there are in fact *two* distinct all-or-none generalizations that can be formulated in terms of these two empirical quantities. The first, already mentioned, is the generalization that the probability of making a *correct response* is constant as long as the subject holds the wrong hypothesis about the concept (i.e., up to the trial of his last error). The second, quite different, is the generalization that the probability of switching to the *correct hypothesis* about the concept does not change over trials (i.e., that the probability is constant that each trial will be the trial of last error).

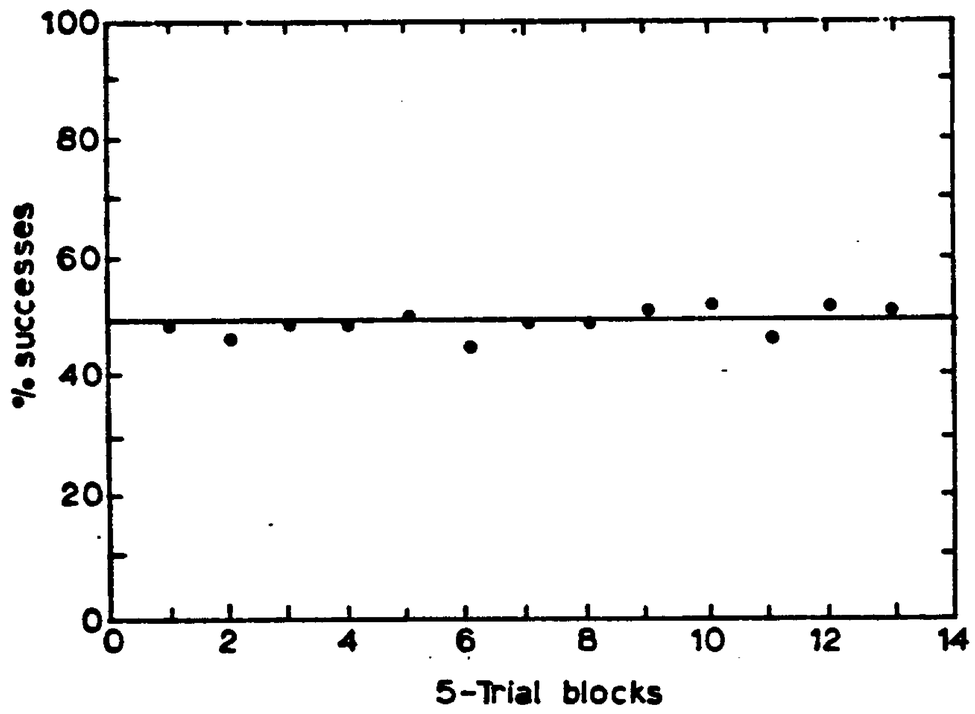

Fig. 5. Concept experiment: percentage of successes prior to the last error (from Bower and Trabasso).

To test the first (correct response) all-or-none generalization, we have one datum from each subject for each trial prior to his last error – a considerable body of data to judge the approximation of the error rate to a constant. To test the second (correct hypothesis) all-or-none generalization, we have only one datum from each subject – the trial on which he made his last error. Hence, trial-to-trial changes in the probability of switching to the right concept are confounded with differences in that probability among subjects. If, for any single subject, this probability increases with trials, the increase is counterbalanced by the fact that the subjects with lowest average probability will tend to learn last. Thus (as Bower and Trabasso are careful to point out) the data to test the second generalization directly are scanty and inadequate.

8. The Bower-Trabasso stochastic model is an explanation of the observed constancy of the error rate. But it is a very bland model, making rather minimal assumptions about the process that is going on. We can pursue the goal of explanation a step further by constructing a more detailed model of the cognitive processes used by subjects in concept attainment, then using this detailed model to subject the theory to further tests. (As Gregg and I have shown in our previous paper on this topic (GREGG and SIMON [1967a]), Bower and Trabasso do, in fact, employ such a process model, but only informally.)

There are two important differences between the summary stochastic model and the more detailed process model. The process model, but not the stochastic model, spells out how the experimenter selects (on a random basis) the successive instances, how the subject responds, and how he selects a new concept when his current one is found wrong. The stochastic model, but not the detailed model, contains two free parameters, one specifying the probability that the subject's response will be (fortuitously) correct when he does not hold the correct concept; the other specifying the probability that he will select the correct concept as his new one when his current concept is found wrong.

The stochastic model and process model can be formalized by stating them in a computer programming language (GREGG and SIMON [1967a]). When this is done, it is found that the stochastic model requires 15 statements – i.e., simple computer instructions – for its formulation, the detailed process model 27. Against this parsimony of the stochastic model must be balanced the fact that that model contains two free numerical parameters, the process model none. Which model is the simpler?

If we apply Popper's criteria of simplicity – the simpler theory being the one that is more highly falsifiable – then the question has a definite answer. The detailed process model is simpler than the stochastic model (see GREGG and SIMON [1967a] pp. 271–272). For, by a straightforward aggregation of variables, the stochastic model, with particular values for the free parameters, can be derived deductively from the process model. Hence, the process model is a special case of the stochastic model. (The process model predicts an error rate of about 0.5 per trial prior to the trial of last error. It also predicts the probability that the last error will occur on a particular trial, but this probability depends on the structure of the stimuli – the number of attributes they possess, and the number of values of each attribute.)

The additional detail incorporated in the process model's assumptions also provides additional opportunities for subjecting the model to empirical

test. The hypotheses held by the subject as to the correct concept do not appear explicitly in the stochastic model; hence data relating to these hypotheses (obtained, say, by asking the subject on each trial what concept he holds, as was done by FELDMAN [1964], or obtained by indirect procedures developed by LEVINE [1966]) cannot be used to test that model, but can be used to test the process model.

If parsimony refers to the brevity with which a theory can be described, then the stochastic model is the more parsimonious (fifteen statements against twenty-seven). But lack of parsimony, so defined, must not be confused with degrees of freedom. We have seen in this case that the less parsimonious theory is the simpler (by POPPER's [1961] criterion), and by far the more falsifiable.

Testing the detailed process theory raises all the problems mentioned earlier with respect to extreme hypotheses. If the error rate on guessing trials deviates from 0.5 should the theory be rejected? How much of a deviation should be tolerated? In how many cases can a subject report he is holding a concept different from that predicted by the theory before we reject the latter? I have given my reasons earlier for thinking that these questions are judgmental, and for concluding that the theory of statistical tests offers no help in answering them. A judgmental answer is that the theory should be rejected only if it turns out to be "radically" wrong. Otherwise, deviations should lead to a search for variables to account for them, and for the "ideal" limiting case in which they would disappear.

Justice Holmes once said: "Logic is not the life of the law". I would paraphrase his aphorism by saying: "Statistics is not the life of science". No existing statistical theory explains what scientists do (or should do) to retroduce, develop, test, and modify scientific theories.

9. Just as statistically significant deviations of data from a generalization should not always, or usually, lead us to abandon the generalization, so we should not be unduly impressed by excellent statistical fits of data to theory. More important than whether the data fit is why they fit – i.e., what components in the theory are critical to the goodness of fit. To answer this question, we must analyse the internal structure of the theory.

For example, under the conditions where all-or-none learning can be expected to take place, the learning trials can generally be divided into two parts: an initial sequence prior to learning, during which the subject can only guess at the correct answer; a terminal sequence, during which the subject knows the correct concept, and makes no new mistakes. Let us suppose that the boundary between these two segments can be detected

(as it can in the concept-learning experiments by the trial on which the last error is made).

Under these conditions, no important conclusions can be drawn about psychological characteristics of the subjects by examining the statistical structure of their responses prior to learning. For the statistics of these responses are simply reflections of the experimenter's randomization of the sequence of stimuli. In one experiment, ESTES [1959], for example, employed three different conditions differing only with respect to the number of alternative responses (2, 4 and 8, respectively) available to the subject (see SIMON [1962]). He found that the relative number of errors per trial made in these three conditions could be represented by the formula, $A(N-1)/N$, where A is a constant and N is the number of alternative responses.

The data on relative numbers of errors fit this formula with great accuracy – a clearcut case of success for an extreme hypothesis of the kind we have been commending in this paper. However, the hypothesis that was being tested was not a generalization about psychology, but a well-known generalization about the laws of probability: that in drawing balls at random from an urn containing white and black balls in the ratio of 1 to $(N-1)$, on the average $(N-1)/N$ of the balls drawn will be black. This is true regardless of whether the subjects themselves, prior to learning, thought they were simply guessing or thought they were responding in selective, patterned ways to the stimuli. By randomizing the sequence of stimuli presented, the experimenter guaranteed the applicability of the laws of probability to the subject's errors, independently of the systematic or "random" character of the subject's behavior.

As I have pointed out elsewhere, a number of other excellent fits of simple generalizations to data can be attributed to the random presentation of stimuli, rather than to characteristics of the subjects (SIMON [1957], SIMON [1962], GREGG and SIMON [1967a]). This does not imply that it is useless to extract the underlying regularities from the data; but we must be careful to provide the regularities with a correct explanation. To do so, we must examine the internal structure of the theories that lead to the successful generalization.

10. Throughout this paper, considerable stress has been placed on the close interaction between hypotheses and data in the building and testing of theories. In most formal theories of induction, particularly those that belong to the genus "hypothetico-deductive" or "H-D", hypotheses spring full-blown from the head of Zeus, then are tested with data that exist,

timelessly and quite independently of the hypotheses[8]. Theories as otherwise divergent as Popper's and Carnap's share this common framework.

It was one of Norwood Hanson's important contributions to challenge this separation of hypothesis from data, and to demonstrate that in the history of science the retroduction of generalizations and explanations from data has been one of the central and crucial processes. In making his point, Hanson was careful not to revert to naive Baconian doctrines of induction. To look at a series of size-rank distributions, approximately log-linear with slopes of minus one; then to conclude that *all* such distributions share these properties, is Baconian. To look at the raw data, and conclude that they can be described adequately by the log-linear function with slope of minus one is not Baconian. It is the latter form of derivation of generalizations from data with which Hanson was primarily concerned, and to which he (following Peirce) applied the name "retroduction".

One of my principal theses here has been that hypotheses retroduced in this way are usually highly plausible, and not highly improbable, as POPPER [1961] would insist. We have already resolved part of the apparent paradox. The "improbability" to which Popper refers is improbability of the very special state of nature described by the empirical generalization, not improbability of the generalization itself. But it remains to understand how the scientist can ever be lucky enough to discover the very special generalizations that describe these a priori improbable (but actual) states of nature.

Fortunately, considerable light has been cast on this question by progress in the past decade in our understanding of the theory of human problem solving (SIMON [1966]). If the scientist had to proceed by searching randomly through the (infinite) space of possible hypotheses, comparing each one with the data until he found one that matched, his task would be hopeless and endless. This he does not need to do. Instead, he extracts information from the data themselves (or the data "cleaned up" to remove some of the noise), and uses this information to construct the hypothesis directly, with a modest amount of search.

Let us consider a concrete example (BANET [1966]). Suppose we are presented with the sequence: $\frac{0}{5}, \frac{4}{3}, \frac{25}{11}, \frac{0}{8}, \ldots$. What simple generalization can we discover to fit this sequence? We note that all the numerators are

---

[8] For a criticism of this view, see SIMON [1955]. In that paper I was concerned specifically with the relative dating of theory and data, and while I still subscribe to the general position set forth there – that this dating is relevant to the corroboration of hypotheses by data – I would want to modify some of my specific conclusions about the form of the relevance, as various paragraphs in the present paper will show.

squares, that the first and third denominators are four less than their numerators, the second and fourth denominators are one less. We notice that the sequence appears to be monotone decreasing, and to approach a limit – perhaps unity. Nine is 3^2, 25 is 5^2. Suppose we number the terms 3, 4, 5, 6. The corresponding squares are 9, 16, 25, 36. Let's multiply numerator and denominator of the second and fourth terms by four, getting: $\frac{9}{5}$, $\frac{16}{12}$, $\frac{25}{21}$, $\frac{36}{32}$,.... Now the empirical generalization is obvious: the general term of the sequence is $n^2/(n^2-4)$. Physicists will recognize this as the well known Balmer series of the hydrogen spectrum, and what we have done is to reconstruct hypothetically part of Balmer's retroduction. (He probably followed a somewhat different path, and we have only considered the last half of his problem of getting from data to generalization, but this partial and somewhat unhistorical example will serve to illustrate our central point. For the actual history, see BANET's [1966] interesting paper.)

However great a feat it was for Balmer to extract his formula from the data, the process he used was certainly not one of generating random hypotheses, then testing them. It is better described as a process of searching for the pattern in the data. It can be shown, for a considerable class of patterns that are of practical importance, in science, in music, and in intelligence tests, that the range of relations the searcher must be prepared to detect is quite small. It may be that these are the sole relations from which the simplicity of nature is built; it may be they are the only relations we are equipped to detect in nature. In either event, most of the patterns that have proved important for science are based, at bottom, on these few simple relations that humans are able to detect.

11. In this paper, I have examined several aspects of the problem of testing theories, and particularly those important theories that take the form of extreme hypotheses. In part, my argument has been aimed at a negative goal – to show that when we look at realistic examples from natural and social science, statistical theory is not of much help in telling us how theories are retroduced or tested.

As an alternative to standard probabilistic and statistical accounts of these matters, I have proposed that we take into account a whole sequence of events:

(1) The enterprise generally begins with empirical data, rather than with a hypothesis out of the blue.

(2) "Striking" features of the data (e.g., that they are linear on a log scale with slope of minus one) provide for a simple generalization that summarizes them – approximately.

(3) We seek for limiting conditions that will improve the approximation by manipulating variables that appear to affect its goodness.

(4) We construct simple mechanisms to explain the simple generalizations – showing that the latter can be deduced from the former.

(5) The explanatory theories generally make predictions that go beyond the simple generalizations in a number of respects, and hence suggest new empirical observations and experiments that allow them to be tested further.

"Testing" theories, as that process is generally conceived, is only one of the minor preoccupations of science. The very process that generates a theory (and particularly a simple generalization) goes a long way toward promising it some measure of validity. For these reasons, histories of science written in terms of the processes that discover patterns in nature would seem closer to the mark than histories that emphasize the search for data to test hypotheses created out of whole cloth.

References

BANET, L., Evolution of the Balmer series, Am. J. Phys. 34 (1966) 496–503.

BOWER, G.H. and T.R. TRABASSO, Concept identification, in: Studies in mathematical psychology, ed. R.C. Atkinson (Stanford, Stanford University Press, 1964) pp. 32–94.

ESTES, W.K., Growth and functions of mathematical models for learning, in: Current trends in psychological theory (Pittsburgh, University of Pittsburgh Press, 1959) pp. 134–151.

FELDMAN, J., Simulation of behavior in the binary choice experiment, in: Computers and thought, eds. E.A. Feigenbaum and J. Feldman (New York, McGraw-Hill, 1964) pp. 329–346.

GOODMAN, N., The test of simplicity, Science 176 (1958) 1064–1069.

GREGG, L.W. and H.A. SIMON, Process models and stochastic theories of simple concept formation, J. Math. Psych. 4 (1967a) 246–276.

GREGG, L.W. and H.A. SIMON, An information processing explanation of one-trial and incremental learning, J. Verbal Learning and Verbal Behavior 6 (1967b) 780–787.

GREGG, L.W., A.P. CHENZOFF and K. LAUGHERY, The effect of rate of presentation, substitution and mode of response in paired-associate learning, Am. J. Psych. 76 (1963) 110–115.

HANSON, R.N., Patterns of discovery (Cambridge, The University Press, 1961).

JEFFREYS, H. and D. WRINCH, On certain fundamental principles of scientifique inquiry, Phil. Magazine 42 (1921) 369–390.

LEVINE, M., Hypothesis behavior by humans during discrimination learning, J. Exper. Psych. 71 (1966) 331–338.

POPPER, K.R., The logic of scientific discovery (New York, Science Editions, 1961).

POSTMAN, L., One-trial learning, in: Verbal behavior and learning, eds. C.F. Cofer and B.S. Musgrave (New York, McGraw-Hill, 1963) pp. 295–321.

ROCK, I., The role of repetition in associative learning, Am. J. Psych. 70 (1957) 186–193.

SAVAGE, L.J., The foundations of statistics (New York, Wiley, 1954).

SIMON, H.A., Prediction and hindsight as confirmatory evidence, Phil. Sci. 22 (1955) 227–230.

SIMON, H.A., On a class of skew distribution functions, Biometrika, 42 (1955) 425–440; reprinted in: Models of man (New York, Wiley, 1957) pp. 145–164.
SIMON, H.A., Amounts of fixation and discovery in maze learning behavior, Psychometrika 22 (1957) 261–268.
SIMON, H.A., A note on mathematical models for learning, Psychometrika 27 (1962) 417–418.
SIMON, H.A., Scientific discovery and the psychology of problem solving, in: Mind and cosmos, ed. R. Colodny (Pittsburgh, University of Pittsburgh Press, 1966) pp. 22–40.
UNDERWOOD, B.J. and G. KEPPEL, One-trial learning, J. Verbal Learning and Verbal Behavior 3 (1964) 385–396.