APPLICATION OF THERMODYNAMICAL METHODS

IN COMMUNICATION THEORY AND IN ECONOMETRICS

by

Benoit MANDELBROT

now at the University of Lille

Research Report for 1956-1957.

SUMMARY
+++++++

Three types of problems were tackled.

A) The theory of the word statistics was resumed, in the spirit of the author's 1951-1954 theory, but under more general conditions. The main postulate is that all the different distinguishable sentences of given coding cost (and, in some cases, given number of words) have equal probabilities. It was shown that the number of words, of cost less than u, may be written as:

$$r(u) = V\left[\exp(b'u - b'C_o) - 1\right]$$

The number V, which was neglected by Shannon, is the inverse of the information per word. It plays the role of the "volume" in the present "thermodynamical" theory.

Several types of "syntax" were considered. It was shown that the formula

$$p(r) = (B-1)\, V^{B-1}\, (r + V)^{-B}$$

holds, whatever the syntax, as soon as the volume V and the temperature 1/B are large. This formula is the counterpart of the formula for perfect gases. The reasons for the success of such a macroscopic theory are shown to be the same as those for the success of the theory of the perfect gas.

The modifications required by the other cases are sketched.

(For details, see a paper published by the Institut de Statistique de l'Université de Paris.)

B) A theory of the law of Pareto was given. The main tool is the definition of the Pareto Lévy variables and sequences. A PL variable is a stable variable, with finite mean, and with maximum positive skewness. A PL process is one in which the joint distributions of the values of u, over several periods of time is a multivariate stable distribution, with finite means and maximum skewness. It is shown that the PL sequences possess all the classical properties as asymptotic properties. The Champernowne transformation is also an asymptotic property of PL sequences. Several types of "explanations" of the PL laws are given.

C) A theory of entropy and a theory of divisibility were added to the author's approach to the foundations of thermodynamics.

INTRODUCTION: A LONG RANGE PROGRAM OF RESEARCH IN STATISTICS.

During any recent period of time, this author's research activity has turned out to overlap on many problems, and to be a rather arbitrary cross-section of the development of a certain single long-range interdisciplinary research program. A report on recent developments may therefore be meaningless, without a previous statement of the general program, and a sketch of its development.

The point of departure was a study of some statistical properties of the natural languages. These have an obvious practical importance, in communication engineering, since the most often transmitted of all messages are texts in natural languages (But the field of the numerical laws of language also offered a most fascinating challenge, to anyone interested in widening - into social science - the field of application of the methods of the natural sciences). The study of language was first carried out as an application of information theory (1951-1954). A set of models was developed, which explained very well the empirical laws of word frequency distribution, discovered empirically by J.B. Estoup and G. K. Zipf. In fact, the theoretical formula, given by the theory, was even a generalization of the law of Estoup and Zipf, and it gave a much better fit to the data. The main feature of the method was that the simplicity and universality of the properties of the words is due to that they are each formed of many letters; this composition "smoothes" the complicated statistical properties of the letters.

In many ways, however, information theory is only an application, to the field of the abstract signals, of some general methods of statistical thermodynamics, a science of bulk phenomena of physics. The theory of language could therefore be considered directly as an application of thermodynamical ideas, to the study of a "bulk" system,

not composed of material particles. A motivation, but no convincing proof yet, was given of the idea, that the improved frequency law is a "linguistic" counterpart of the structure of the perfect gases of physics. However, of course, if such an application of thermodynamics were not to be shocking, one should refer to a set of foundations, which does not use any too particular kinetic consideration; in fact, a thermodynamic theory of language should better be accompanied by a special study of the foundations themselves. This would be a new motivation for the study of foundations: to distinguish the part of thermodynamics, which is so general that it could also conceivably apply to systems more general than sets of molecules. In particular, the information-theoretical interpretation of the model of language could be preserved all through the theory, and this new use of thermodynamical methods could contribute to their understanding in general contexts.- However, this author's first study of the foundations (1952) was still too much based upon considerations of the relationship between entropy and information, and it was no more successful than similar contemporary studies.

But the "uncritical" application of thermodynamical methods was again successful, in the study (in 1954-1955) of other problems of great practical (and conceptual) interest: the problem of the classification of the items of a finite set, through successive dichotomies (and through intermediate categories), and the connected problem of information retrieval. The practical importance lies here in the fact that the design of "good files" is well known to be one of the stumbling blocks in the efficient design of some very large organizations. (The theoretical interest is also clear: for example, in the case of biological taxonomies, the structure of the classifications, - which are already available in this case - was widely believed to provide certain information about the organization of the living beings, for example about the laws of evolution). The 1954-1955 study was limited to the examination of J.C. Willis's law, giving the distribution of the species among genera. This law was explained on the basis of "thermodynamical" considerations, (without reference to evolution).

There are many differences between the laws of Estoup-Zipf and of Willis (and some superficial resemblances: they can both be written in the form $y = P x^{-a}$, with appropriate interpretations of x and of y, and with appropriate values for P and a. But this hardly goes any further). But there is also a deep common feature: when the "energy", u, is properly defined, the number, s(u) of configurations of the system, having each the given energy u, increases very much more rapidly than is ever the case in physics. As to the "phase space", it cannot be an euclidean space, as in physics, but it is some simple function space. Finally, the partition or generating function: $g(b) = \sum s(u) \exp(-bu)$ has now a singularity in the right half-plane, whereas in physics, the origin is the singularity of largest real value. One consequence is that the distribution of sums of very many identical and independent random variables is no longer necessarily gaussian. It may be one of the non-gaussian stable probability distributions of Paul Lévy (expressions which were, until recently, considered as entirely pathological).

The part of thermodynamics, which can be generalized, should therefore exclude any assumption, which would make the phase space necessarily euclidean, or which would make the sum of many components, necessarily gaussian. A new axiomatics for this part was developed in 1955-1956, which is both statistical and "phenomenological", in that it never involves the atoms as "hidden variables". (In some ways, this approach turned out to be equivalent to an old - and undeservedly neglected - approach, used by L. Szilard in 1925). The new foundations are not yet sufficient to cover all the non-physical applications of thermodynamics. But they turned out to have a most interesting feature: the main tool which they use is the mathematical theory of statistical estimation of parameters, which is also the main tool of the theory of the detection of weak signals in noise. The theory of the foundations of noise can thus be based on the same

principles, as the theory of elimination of noise, in communication. Such a unity is quite desirable, and it was planned in 1956 mostly to go ahead with the development of such an approach to the foundations, and to let non-physical applications aside, for a while.

These plans could not be followed, and, finally, more striking progress was made in the study of the applications (§). It appears that the author's activity has by now been durably oriented along the lines of a long range program, having two aspects: the study of physical and non-physical applications of thermodynamics, and the study of the "universal" foundations. It seems that the progress in either aspect is very much conditioned by that in the other, and it is planned to pursue both fairly simultaneously, in the near future.

The topics most studied in 1956-1957 were

I. The problem of word frequencies, which was resumed after several years. A research paper on this topic has been published under the title "Théorie mathématique de la loi de Zipf"; publisher: Institut de Statistique de l'Université de Paris. An English translation, with complements, is being kindly made by members of the IBM Laboratories.

II. The problem of the law of Pareto, relative to income distribution, a new topic, providing a different type of generalization of thermodynamics. A research paper on this topic is far advanced.

III. The problem of foundations. A research paper on this topic is in progress.

Chapters I, II, and III of the following account are each devoted to one of the above items. Chapters I and II contain furthur considerations on the relationship of these applications to other problems. Chapter II includes some results obtained, during a summer stay at Cornell University, on Contract with the ONR Logistics Project of Columbia University.

(§) To the progress made in the theory of foundations, one should however add the efforts of prof. L. Tisza, and of Father Quay, both of M.I.T., who became interested in this author's approach, after the conferences he gave at MIT in 1956.

CHAPTER I. THE NEW GENERALIZED FORM OF THE "THERMODYNAMICAL" THEORY OF ZIPF'S MICROSCOPIC LAW OF WORD FREQUENCIES.

1.1 Statement of the law of Estoup and Zipf. Everyone has a strong feeling of the very great complication of the statistical properties of continuous samples of natural texts, when these samples are considered as either sequences of letters or of phonemes. Besides, these statistical properties clearly differ very much from language to language and from author to author. One could think then that even more complicated statistical properties would be observed for the higher-scale units, such as the words, which are each made out of many letters. However, this is not the case. It has been observed, at least since J.B. Estoup(1916) and it is fairly well known since G.K. Zipf, that the statistical distribution of the words is essentially independent from the language and from the author considered. "Essentially" means in this case that the distribution depends only upon very few (two) numerical parameters. Of course, meaning, etc... could not be taken account of, in such a type of law. It turns out, that the best way of writing the experimental result starts by a "neutral" relabeling of all the words: one uses as rank, r, the position in which they are found, in the ordering of all the words of a large sample, by decreasing frequency. The empirical data turn then, in most cases, to be excellently represented by the following law, which is a generalization of Estoup and Zipf's original expression

$$p(r) = (B-1)\, V^{B-1} \quad (r \blacklozenge V)^{-B}$$

where p(r) is the probability of the word of rank r, and where B and V are the two parameters. For large values of r,

$$\log p(r) = \log (B-1)\, V^{B-1} \quad -B \log r$$

That is: on bilogarithmic paper, log p(r) (as a function of log r) is a straight line. For small values of r, the empirical points fall on both sides of the above concave curve, with deviations on

both sides adding to zero.

In § 1.3, it will be shown that the formula can be entirely explained, by considering long sequences of words as thermodynamic systems, and by applying to them the physical principle, that all the configurations of a microcanonical system are equiprobable. In fact, not only the formula can be explained by the thermodynamic theory, but it was itself obtained by this theory, and turned out to be superior to the less general formula, first suggested by the experimenters. The theory will also provide for several possible deviations from the above formula.

1.2. On general thermodynamic systems, and on their classification.

1.2.1 Equirepartition of probability among the distinguishable configurations of given energy. Assume the following principle:

To any "thermodynamic" system, one may attach a quantity, u, called its "energy". Let dr (u) be the number of distinguishable configurations of the system, such that their energy U is such that $u < u < u+du$. When u is fixed and known, all these configurations have equal probabilities.

Of course, the first difficulty of such a theory, in the case of non-physical systems, will be to define the "energy" correctly. In the case of physical systems, there is no such difficulty, and the function r(u) may be identified to the volume of the shell of energy u, in the phase space of the system. This phase space is euclidean; in the course of the argument, the number of its dimensions will tend to infinity, but all the arguments are made on finite dimensional spaces. For the perfect Maxwell Boltzmann gas, formed of N molecules,

$$r(u) = K\ V\ u^{3N/2}$$

Therefore,

$$g(b) = \int \exp(-bu)\ dr(u) = K'\ V\ b^{-3N/2}$$

For other physical systems, r(u) and g(b) may take other forms. But it is always assumed very early, in physical arguments, that g(b) has no singularity in the positive half plane (real part of $b \geq 0$).

Besides, one assumes that $r(\infty) = \infty$ (this is certainly true, as soon as u is unbounded). Therefore, there is always a singularity for $b = 0$.

As pointed out by J.W. Gibbs, these conditions are not required by the laws of mechanics and of non-statistical thermodynamics. However, they destroy the generality of statistical thermodynamics, at an early stage. One could therefore conceive of at least two types of generalizations.

A) g(b) could be regular for $b = 0$; this requires $r(\infty) < \infty$, and u must be bounded if r(u) is integer-valued. In that case, g(b) will be everywhere regular. Such systems have recently been considered in the study of negative temperatures; they had also occurred in the study of some quite exceptional linguistic systems (see § 1.3.5).

B) g(b) may have singularities in the positive half-plane; since $dr(u) \geq 0$, one of the singularities, having maximum real value, must be real; let it be called "leading singularity" or b'. Physical systems for which $b' > 0$ could absorb or abandon an infinite amount of heat, without exceeding a certain temperature $1/b'$, and without work; they are therefore completely foreign to our common experience of nature, as Gibbs found out for all the examples which he investigated

One exception is the infinite thermostat, for which $r(u) = K \exp(b$ but such a thermostat only exists at a single temperature $1/b'$.

However, the physical intuition becomes worthless in the non-physical applications which are the object of the present theory.

One consistent feature of the non-physical applications of thermodynamics, involved in the present research program, is that their phase space is not euclidean, but is some simple infinitely dimensional function space. The function r(u) then increases so rapidly, that g(b) has singularities in the right half plane. The theory of such systems sometime preserves, and sometime distorts, the usual features of gas theory, in a most fascinating pattern.

1.2.2 Possibility of a macroscopic description, when b is near the leading singularity b'. The second step of thermostatistics is to introduce the canonical distribution:

$dF(u/b) = g^{-1}(b) \exp(-bu)\, dr(u)$, where $F = Pr[U < u]$

In the physical case, $b' = 0$, this may be obtained as being either the most probable, or the average, distribution of energy for a small part of a large microcanonical system. It may be seen that this identity no longer necessarily holds, if $b' \neq 0$; but the canonical formula keeps its importance. It may be noted that the properties of thermodynamic systems of molecules are simplest, when the temperature is high, that is, b near 0. Then, the low energy levels rarely occur, and the form of r(u) for u small is not important. But the form of r(u) for large u only depends upon the behavior of g(b) near its leading singularity b', so that the very possibility of a simple macro description of a system is linked to the possibility of approximating g(b) by its principal value near the singularity. The macro description blows up, when the temperature becomes small.

One can proceed likewise, for non-physical thermodynamic systems. If the temperature 1/b is extremely high, that is, if b is very near its minimum b', one can construct a macro theory, with very few parameters, by only considering the behavior of g(b) near b'. Furthermore, all the theories which one may encounter may be classified, on the only basis of the mathematical character of the leading singularity.

To "explain" why a certain set of empirical data seems to correspond to a certain r(u) or g(b), one may, in the first approximation, simply justify the behavior of g(b) near $b = b'$, that is, certain broad features of the phase space. All the present applications are still in this first approximation stage.

The only singularities considered so far are poles at $b' \neq 0$, and branching points. In both cases, the nature of the singularity of the sum of N systems is easily deduced from that of single systems.

(In the Pareto case, no good explanation was yet found of the

singularity required by the data. Therefore, the Pareto theory will be presented on different grounds).

The success of the thermodynamic approach will be translated in all cases by the fact that the concrete nature of the phenomena concerned will become irrelevant, once the g(b) function is derived. Moreover, one finds that b is consistently close to b', or even equal to b'. One should not yet look for very deep reasons for this fact, because, for large b, the predictions of thermodynamics become so involved, that the verification of the principles involved would be most difficult at best. Besides, the striking fact (at this stage) is that a macroscopic description, and theory, apply at all: it is quite plausible that the present limited and scattered set of numerical data is heavily biased in favor of the cases where the results are simple. When any particular field is studied exhaustively, one is bound to find an increasing number of exceptions, and these may become the more interesting part of the whole approach.

1.2.3. Absolute temperature. Let g(b) be only defined for $b > b'$ or for $b \geqslant b'$. The absolute temperature is defined in physics as $1/b$; its scale is quite arbitrary, unless it is chosen so as to make Boltzmann's constant k equal to 1. In the cases where $b' \neq 0$, the choice of a scale of temperature is much more straightforward: one may define

$$1/B = b' / b \quad (B = \text{capital beta})$$

Such non-physical systems can thus only be defined for temperatures less than 1. (The above definition is also equivalent to a choice of the unit of u, so that b' becomes = 1.)

(One finds however that some of the poles of the physical temperature are played by $b'/(b-b')$, and not by b'/b.)

1.2.4 On thermodynamics at unit temperature, and on stable, non gaussian distributions. In the physical case, $g(b') = g(0) = \infty$. Therefore, the canonical formula $dF(u/b) = g^{-1}(b')\ \exp(-bu)\ dr(u)$ gives $F = 0$ for all u, which is without any interest. The state

$b = b'$ remains without interest, as long as the leading singularity is a pole. But, in the case of other types of leading singularity, $g(b')$ may be finite; then, the canonical formula is a genuine probability distribution, even at the singular point.

Consider now the sum of a large number of systems, independent and identical, and such that $g(b') \neq \infty$. Write $dr'(u) = \exp(-b'u)\, dr(u)$. (This function of u must decrease rapidly enough, so that $r'(\infty) = g(b') \neq \infty$). It is no longer necessarily possible to apply here the local central limit theorem, which Khinchin uses in the physical case. Even if the sum tends to a limit (when properly normed), the limit may be any one of Paul Lévy's stable probability distributions, that is, distributions having a characteristic function of the form:

$$\exp\left[C\ |t|^{\alpha} \left\{ 1 + \frac{it}{|t|} \beta\ \mathrm{tg}\ \frac{\alpha\pi}{2} \right\} \right]$$

where $0 < \alpha \leq 2$, and $|\beta| \leq 1$. We shall return to the case in § II. The most striking property of such variables is that if, in this case, a large microcanonical system is divided into two parts having the same probability distribution, each part is no longer probably equal to $\frac{1}{2}$ of the whole, plus or minus a small fluctuation; on the contrary, one part will most probably take up most of the whole, leaving only a small remains to the other; only, one does not know in advance which part will be the greater.

1.3 On the theory of the Estoup-Zipf law, relative to the word; a system, the structure function of which is an exponential.

(This § will be sketched rapidly, since a complete account has appeared in the author's "Théorie mathématique de la loi de Zipf", Institut de Statistique de l'Université de Paris, June 1957).

1.3.1 Introduction. The principles and methods of statistical thermodynamics (classical and quantum) were applied, with some minor and some major modifications, to the study of the statistical structure of a class of objects, the words. A number of known properties of the words of natural languages were thus explained, by being reduced to criteria, that are both simple and robust, in

that the final results of the theory are not modified by slight modifications of the criteria, and do not depend upon some arbitrary assumptions, which must be made at the beginning of the theory. Several conjectures were made, concerning aspects of natural language that have not yet been explored experimentally.

1.3.2. The cost of transmission as "energy". Morphology. Number of different words of given cost volume. The macroscopic systems of the theory will be the words, sequences of lesser, a-tomic units, called letters (in contrast to the words, the letters will not be identified to the letters of natural language; but discrete units much lesser than the words must exist). In a continuing sequence of letters, a natural word will be any string, contained between two successive occurrences of a certain improper letter, called the space. The morphology of language will always be taken to be of Maxwell Boltzmann type, that is: all permutations of different letters will be different words (some may never occur actually: this should be expressed here by a zero probability of occurrence). Therefore, the number of different words, in which the letter L_m occurs $n(m)$ times, will be $\prod n(m)\ !\ /\left[\sum n(m)\right]!$ (whereas, in physics, permutations should not be distinguished, so that one must somehow suppress the term $\left[\sum n(m)\right]\ !$

The energy attached to each letter will be the cost of transmitting it over a certain channel; this cost will be assumed to depend at most upon the letter itself, and the one which preceeds it in a continuing string. The cost of a word will be the sum of the costs of all the letters which compose it.

It is shown that the generating function $g(b)$ takes the following form, when the cost, C_m, depends only upon the letter, L_m (L_o being the space)

$$g(b) = \exp(-bC_o)\left[1 - \sum \exp(-bC_m)\right]^{-1}$$

As to $r(u)$, it varies by integral jumps, since the words are discrete. In the case where the C_m are not all multiples of some common unit, the behavior of $r(u)$ may be very complicated. However, it is then

possible to approximate the distribution of energies of words, by a continuous distribution, with a function $r^{\circ}(u)$ replacing $r(u)$. In any case, such a function $r^{\circ}(u)$ must be such that $g^{\circ}(b) = \int \exp(-bu)\, dr^{\circ}(u)$ be an approximation to $g(b)$. One may take

$$g(b) = \exp(-bC_o) \left[C_m \exp(-b'C_m) \right]^{-1} (b-b')^{-1}$$

This gives

$$r(u) = V \left[\exp\left(b'(u-C_o)\right) - 1 \right]$$

where

$$V = \left[b'C_m \exp(-b'C_m) \right]^{-1}$$

is the only remaining influence of the initially postulated set of letter costs.

The above formula will not fail to be compared with the formula relative to the number of configurations of a perfect gas (contained in the volume V) which have an energy less than u. In the quantum theory of perfect gases, these configurations are discrete, and they depend upon the shape of the reservior; but, for a large volume, one may introduce an approximate continuous distribution of levels, and the dependence upon the shape reduces to a simple proportionality to the volume of the reservoir. In other terms, it seems that V will play the role of the volume. (However, the universal function which multiplies V is no longer a power of u, but an exponential; this change is due to the fact that one keeps the factor $\left[\sum n(m)\right]!$, and to the fact that the number of letters in a word is not fixed, which recalls the case of photons; however, the present system is a Maxwell Boltzmann system, and not a Bose Einstein system, like the photons). The interpretation of V as volume is linked to a most interesting phenomenon, a counterpart of Gibbs's paradox of thermodynamics, for which we must refer to p. 26 of our special report.

1.3.3. Syntax. Temperature. Consider now the sentences, which are sequences of words. Two models of syntax were studied. In the Maxwell Boltzmann syntax, all the permutations of different words are considered as being correct and distinct sentences. In the

Bose Einstein syntax, the permutation will not change the sentence; no restriction will be set a priori on the length of the Bose Einstein sentences, but it will be shown that, a posteriori, the length of most sentences will be "essentially" bounded. The number of words in a sentence may be fixed in advance (uniform sentences) or be determined by successive occurrences of some special word: the "full stop" (natural sentences). It may be noted, that these definitions have something quite unusual from the viewpoint of physics. (In natural M.B. sentences, the M.B. count is associated with variable length, whereas in physics, variable length is found only with the B.E. count. In the B.E. syntax, the words, which are built according to the M.B. count, are composed according to the B.E. count; in physics, only the opposite is found.)

The basic postulate is still that all sentences of given energy (the number of words in the sentence may be fixed in advance, or not) have the same probability of occurrence. If so, it is no longer true that the most probable and the average frequencies of words in sentences are necessarily equal. But in the case of the uniform sentences, the two specifications are identical, and lead to

$$\bar{N}(r) = \left[Z^{-1} \exp(bu_r) - C \right]^{-1} \qquad \begin{cases} C = 0 \text{ in the M.B. case} \\ C = 1 \text{ in the B.E. case} \end{cases}$$

where r is the index of the word in some arbitrary list of words. The values of Z and of b are determined by the relationships:

$$N = \int dr(y) \left[Z^{-1} \exp(by) - C \right]^{-1} \sim V \int \exp(x)\, dx \left[Z'^{-1} \exp(Bx) - C \right]^{-1}$$

$$b'u = \int dr(y)\, yb' \left[Z \exp(by) - C \right]^{-1} \sim V \int \exp(x)\, dx \left[Z' \exp(Bx) - C \right]$$

where $Z' = Z \exp(bC_o)$ and $B = b/b'$

Thus, Z' and B are functions of the ratios N/V and b'u/V. The influence of the B.E. syntax will depend upon the effect of the value of C in the above formulae, which in turn will depend upon V. The role of V will therefore again be the same as the role of volume in physics.

1.3.4 The perfect word. In particular, if B is very close to one (high temperature) the syntax has no influence upon the ratios N(r)/N, and, in all cases, thermodynamics predicts that

$$p(r) = N(r)/N = \exp(-bu_r) \left[\int \exp(-bu)\, dr(u) \right]^{-1}$$

There is no way, however, of checking the predictions of thermodynamics, as long as they are presented in such a form, because u is not something which can be easily measured. There is undoubtedly some weakness in a theory starting from the identification to energy of a non-measurable quantity, but this weakness is not harmful, after all. First of all, experimentally, one can short circuit u, in the verification of the predictions of thermodynamics. This is because u is a well-defined function of the number of words of energy less than u, and this number is itself equal to the number of words more frequent than a word of energy u. Suppose that the above arbitrary index r is chosen to be the rank of a word, in the ordering of all words by decreasing frequencies; it can then be identified to the r of the formula for r(u).

Further,

$$r(u) \sim \overset{\circ}{r}(u) = V \left[\exp\ (b'\ (u-C_o)\ \ -1 \right]$$

Eliminating u between r (u) and p(r), one obtains, if V is large,

$$p(r) \sim \overset{o}{p}(r) = \ \ (B-1)\ V^{B-1}\ (r+V)^{-B}$$

Let us return to the conditions of validity of this prediction. First, one must have a B close to 1 (high temperature) and a large V (large volume). But, besides, the formula for r (u) is an approximation valid for large u, based upon an approximation of g(b) near b'. Therefore, the formula $p^o(r)$ is itself strictly valid only for large r. For small values of r, p(r) will differ from $p^o(r)$, but $p^o(r)$ was constructed in such a way, that the positive and negative differences will add to zero. These differences are due to the details of the letter costs, of which the macroscopic theory cannot take any account. To these differences, may also be added the chance fluctuations, inevitable in small samples.

As mentioned, the formula $p^o(r)$ does in fact give an excellent prediction of empirical facts. Besides, its conditions of validity are now clearly shown to be formally the same as those of the validity

of the theory of the perfect gas. Hence the proposed terminology: "perfect words".

In the memoir , the properties of perfect words are studied in more detail.

1.3.5. Other states of the word. The word statistics will differ from that of the perfect word if the volume or/and the temperature are small, or if there is some artificial bound on u, so that the temperature can go beyond the value 1.

If the syntax is M.B., but the volume is very small, there is no simple macroscopic prediction of p(r) for small r. The curve of log p(r) as a function of log r may present all kinds of behaviors, more complicated than a simple concavity. But for large r, the formula for p(r) remains valid. That is:

$$\Pr(\ r\ /\ \text{knowing that } r > 5, \text{ say}) = r^{-B\left[\sum_{6}^{\infty} x^{-B}\right]-1}$$

If the syntax is B.E., and V is small, there are two types of effects. First, some modifications of the curve p(r) for small r, which cannot be described in macroscopic terms. Second, and more important, it is found that the average number of words in a sentence, other than the most frequent word, is bounded. Therefore, when the length of a sentence is made to increase without bound, one can add nothing but new repetitions of the most frequent word. (Except in the always present small percentage of exceptional cases.) The bound on the "essential" number of words in a sentence is of the same order of magnitude as the average number of letters in a word. Assume that the number of repetitions of the most frequent word can be disregarded This result means, that the Bose Einstein syntax will never need to be applied to sentences, which are so long, that it would be obviously absurd to postulate that the words in this sentence can be permuted without change of meaning. Thus, the B.E. syntax contains the limits to its own applicability. This phenomenon is the linguistic counterpart of the Bose Einstein condensation of gases. Let the number of words exceed the bound above; the excess words, all identical to the most frequent one, will be the counterpart of the part of a gas

which is liquified when too many molecules are squeezed into a given volume.

Let finally the number of words, and therefore u, be bounded. Then B may become less than one. Then V becomes quite unimportant, and its role is played by the total number of words, R. One finds:

$$p(r) = (B-1)\, R^{B-1}\, r^{-B}$$

Such statistics seem to be found in cases such as that of modern Hebrew, etc... They resemble the physical systems of negative temperature.

CHAPTER II. ON PARETO'S MACROSCOPIC LAW OF INCOME DISTRIBUTION: A NEW INTERPOLATION AND THEORY: PARETO-LEVY STATIONARY STABLE PROCESSES.

2.1 Empirical data and classical theories.

2.1.1. Different forms of the law of Pareto. V. Pareto had attempted to represent mathematically the inequality of distribution of income among the income recipients.

Strong law of Pareto. Let an income recipient be chosen at random from a large population. The probability that his income U over some time interval will exceed the value u, is

$$1 - F(u) = (u/u^{\circ})^{-a}$$

Weak law of Pareto. (or: Doeblin's condition) This strong law is surely incorrect for small u. One could only assert that it holds for large positve u, where one should find, in bilogarithmic coordinates:

$$\log\{1-F(u)\} = a \log u^{\circ} - a \log u.$$

The law of Pareto has been discovered, and is still being studied, by observation of such bilogarithmic graphs. But, in fact, the best that can be drawn from the experimental data is the weaker assertion that a can be chosen so that:

$$P(u) = \log\{1-F(u)\} = a \log u^{0}$$

depends only little upon u. That is, whichever w, there exists a such that

$$\log P\{\exp(\log u + \log w)\} - \log P\{\exp(\log u)\} \to 0, \text{ as } u \to \infty.$$

That is, for all h, $P(hu)/P(u) \to 1$, as $u \to \infty$. As to $Pr(U < -u)$, it decreases very much more rapidly, as $u \to \infty$.

This only reasonable inference from experience turns out to be entirely identical to the necessary and sufficient condition of Doeblin's theorem of the theory of probability.

There is some question, about the range of variation of a. In a few cases, but very doubtful ones, it seems that $a > 2$. But there are such deep differences, between the case where $a > 2$, and that when $a < 2$, that we shall provisionally assume that $1 < a < 2$.

Then, the mean value of U is finite, since:

$$E(U) = \int u\, dF(u) = \int \{1-F(u)\}\, du < \infty$$

But the mean deviation from the mean is infinite, since:

$$\sigma^2(U) + E(U)^2 = \int u^2\, dF(u) = 2\int u\{1-F(u)\}\, du = \infty$$

Pareto's data and laws refer to incomes over fixed time intervals. It is clear, however, that the covariances of incomes over distinct time intervals will a fortiori be infinite.

As a consequence, the usual theory of second-order stationary time series will certainly not apply to income time series.

2.1.2. Models of the Law of Pareto, as expressed in terms of $v = \log u$. It seems that, so far, all the attempts to "explain" the law of Pareto were based on the replacement of u, by Weber-Fechner's logarithm:

$$v = \log u,$$

as the fundamental variable. The law of Pareto then takes the form:

$$Pr(V > v) = \exp\{-a(v-v^{0})\}$$

This exponential distribution is most common in physics, and it turns out that the known theories of the law of Pareto are simply transliterations of known physical theories. These

transliterations are a priori fairly reasonable as such, in the frame of our generalization of thermodynamics; but, a) there is no clear way of justifying the replacement of u by v; b) there is no way of justifying the restriction of the values of a; c) let the time interval, over which the income is studied, tend to zero: the classical theories do not lead to a reasonable interpolation of the curve of variation of income. If, however, it were decided to disregard these fundamental objections, there would be no reason to prefer one "classical" theory to another, since they are as equivalent, as different approaches to the theory of themodynamic equilibrium.

Let v be referred to as "income bracket".

The Pareto distribution, as the most probable, or average, distribution of log u. The formula for $Pr(V > v)$ is simply the canonical formula of thermodynamics, in the case where there is a single "configuration" per income bracket: $r(u) = u$ $dr(u) = du$. That is, if the average bracket $\int dp(v)\, v$ is given, the most probable distribution $p(v)$ is obtained by maximizing $-\int dp(v) \log p(v)$, and it is the Pareto distribution. This argument implies that the income brackets can be added. If so, one may, alternatively, consider all the partitions of a given sum of brackets among N individuals. If these partitions have equal probabilities, the mean frequency of the bracket v is also given by the Pareto distribution.

(An approach of M. Castellani is a slight variant of these arguments.)

Champernowne transformations. Let the income bracket be quantized, the k-th bracket including incomes such that $k\hat{v} < v < (k+1)\hat{v}$. Champernowne conjectures that, when h and k are both large, the probability that the income transfer from the k-th to the h-th bracket is only a function of h-k.

Mathematically, this is the same as conjecturing that the income bracket perform a "random walk", as time proceeds. The simplest random walk is one, in which there is the

probability p, that the bracket increase by one unity, a probability q that it decrease, also by unity, and the probability 1 - p-q, that it stay unchanged. It is known that a random walk is a discrete form of diffusion; that is, of the heat motion of particles in a reservoir of uniform temperature. Such a representation of economic interactions by heat motion is a quite natural one, and is quite familiar; and one should at least try to see its consequences. Champernowne states that, under some additional conditions, the Pareto distribution is the only steady one, under diffusion.

(If there were no additional conditions, it is clear that the income brackets would diffuse to $+\infty$ or $-\infty$, and there could be no steady state.)

The simplest conditions, in physics, under which there exists a steady distribution, are those where the diffusion has a downward average trend, and where there exists a lower floor, under which the particles cannot go, and which bounces or reflects them back. These are precisely the additional conditions chosen by Champernowne. Then, in the case of the simple random walk, the proof of his assertion becomes immediate! In that case,

$$Pr(k,t+1) = p\ Pr(k-1,t) + q\ Pr(k+1,t) + (1-p-2)\ Pr(k,t)$$

except for k=1, since the q% of the people in bracket 1, who would go down, are taken to be "reflected" back into k=1 by some benevolent power. Thus,

$$Pr(1,t+1) = q\ Pr(1,t) + q\ Pr(2,t) + (1-p-2)\ Pr(1,k)$$

This has a steady solution (see Feller, p.):

$$Pr(k) = (1-p/q)\ (p/q)^{k-1}\ ;\ Pr(V \geq v) = \exp\left\{-V\ \frac{\log (q/p)}{v}\right\} = u^{a} u^{-a}$$

which is the Pareto distribution, with:

$$a = \log(q/p)(\hat{v})^{-1}$$

Clearly, the same result would be found, if the reflection were not immediate, but if the person could stay in bracket 1 for a random interval of time, having an exponential distribution. Or else, the bracket 1 could absorb the persons, (that is: they could be ruined for life, if they get there) on the condition

that this be compensated by a uniform flow of new persons into the lowest income bracket.

Now, look only at the behavior at infinity. If the reflecting barrier were placed at $v_o \neq 0$, the coefficient $U^{\rho a} = 1$ would be replaced by $\exp(-av^{\rho})$. Thus, the coefficient a depends only upon the transition probabilities, but the coefficient U^{ρ} depends on the position of the reflecting barrier.

Finally, the steady state under the Champernowne transformations is also the most probable state.

Reflecting layers, and the log-normal distribution. One must now give an account of the fact, that the strong law of Pareto does not hold for small v. There is no lower floor for incomes, but a certain most probable value is followed by a rapid decrease. To explain this on the basis of a random walk, one must assume that log (p/q) starts decreasing for small v, and even becomes negative for very small v (one would like to take account of the possible, but rare, negative incomes u; but this is prohibited by the fact that one works with v = log u). In other terms, log Pr(v) was a linear function of v, for large v; but for intermediate and small v, it will be a function concave towards large values of log Pr(v). As a first approximation, this concavity could be represented by a parabola, which gives,

$$\Pr(v) = K \exp\left(- K' (\log u - \log u_{max})^2 \right).$$

This is the lognormal distribution, for which it has been repeatedly claimed, that it represents the income data in the intermediate range. The relationship of the above explanation, to the usual limit arguments, has been studied in detail.

Criticism of the failure to distinguish between different values of a. The expected value of v decreases, whichever a. But u itself may either decrease or increase on the average. One has:

$$E(U,\ t+1;\ u,t) = \left\{ p\exp(\hat{v}) + q\exp(-\hat{v}) \right\} u$$
$$E(U-u) = \left\{ \exp(\hat{v}) - 1 \right\} \left\{ 1 - \frac{q}{p} \exp(-\hat{v}) \right\} pu$$

whereas:

$$a = \log (q/p) \ (\hat{v})^{-1}$$

Therefore, U decreases on the average, if $a > 1$.

U increases on the average, if $a < 1$.

There is no way of distinguishing such different behaviors, in the frame of the Chempernowne theory.

2.2. Definitions of Pareto-Lévy variables and processes.

In contrast to all previous theories, no unjustifiable functional transformation of u will be placed at the beginning of the present new theory of the Pareto law.

This theory will be based upon a certain interpolation of the weak law of Pareto, for all values of u; the Pareto-Lévy (PL) law. The interpolation could be considered as a conjecture about the actual behavior of the income data, for the intermediate and negative u; such a conjecture could be verified only when the law of PL is tabulated. But the interpolation could also be considered as only a mathematical device, easy to handle, which surely preserves the known facts.

A Pareto-Lévy random variable will be a variable having the following characteristic function:

$$\varphi(t) = \int \exp(itu) dF(u) = \exp\left\{ -c\,|t|^{a} \left[1 - \frac{it}{|t|} \operatorname{tg} \frac{a\pi}{2} \right] - it\,d \right\}$$

where $1 < a < 2$.

Such variables belong to Paul Lévy's class of non-gaussian stable variables, and have been studied in great detail in pure probability theory. It is well known that the probability of positive values of u satisfies the weak law of Pareto, and that the probability of negative u decreases very rapidly. It will be shown that the stable distributions are generalizations of the gaussian. Recall then that for variables with finite second moment, one may construct a gaussian variable with the same first

and second moments. Similarly, if the second moment is infinite, but the weak Pareto law is satisfied, one may replace u by a PL variable having same mean value and moments of order a - 0.

One also knows that a gaussian random process is one for which the joint distribution of the values of the random function, at different instants of time, is a multivariate gaussian. Similarly, a sequence of variables will be called Pareto-Lévy, if the joint distribution of values at different instants of time is a multivariate P-L distribution.

In the bivariate case, a PL distribution of the vector $\bar{U} = (u(t'),u(t''))$ is such that:

$$\bar{U} = \int_0^{\pi/2} u(t)\ \bar{I}\ dG(z)^{\ 1/a}$$

where u(t) is a PL variable, I the unit vector in the direction z from the u(t') axis, and G(z) is a certain scale function. Thus, bivariate variables are sums of PL variables distributed in all directions, from that of u(t') to that of u(t''). Similar definitions apply for trivariate PL distributions.

It will be shown that in the Markovian case, where u(t'') depends only upon u(t'), the dependence expressed by the PL bivariate distribution is asymptotically reduced to the random walk, considered by Champernowne.

2.3. Properties of Pareto Lévy variables. Let u_i be identical and independent variables, satisfying the weak Pareto condition. Consider the expression:

$$u = \lim \left\{ n^{\ 1/a} \left[u_i - E(U_i) \right] \right\}$$

By a classical theorem of Doeblin, u is a PL variable. Conversely, in order that u be PL, it is necessary that the u_i satisfy the weak Pareto law. (This necessary and sufficient condition was the main motivation for the form chosen for the weak Pareto law.) As a consequence, if u' and u'' are identical PL variables,

$$u = (c'u' + c''u'')\ c^{-1}$$

will be the same PL variable, if $c^{1/a} = c'^{1/a} + c''^{1/a}$. This is called the property of stability under addition of the set of variables of the form cu.

There exist stable variables, other than the PL ones, and each of these is a possible limit of normed sum of identical and independent variables. However, among the stable distributions, the PL is the only one for which the mean value is finite, and which is very skew, in the sense that, for large u, the probability of -u is of a much smaller order of magnitude than that of u itself. Equivalent forms of this condition will be given later.

These theorems, quite classical in pure probability theory, are in contradiction with the still very widespread belief, that a limit of normed sums of random variables is necessarily gaussian. As applied to the income data, this belief seemed to be in contradiction with the law of Pareto. To preserve the idea of additivity, some were even prepared to assume that one should not add the incomes, but some function of incomes; but the only function of u ever seriously considered, v = log u, does not lead to the weak law of Pareto, either.)

In fact, however, it is seen that no scale transformation is necessary or bearable. To determine the PL distribution among all distributions, it is sufficient to note that they are the only ones which are limits of normed sums of identical and independent random variables, and which satisfy some additional conditions, which eliminate the gaussian. So far, it was mentioned that extreme skewness is sufficient; other equivalent conditions will be mentioned shortly.

This explanation can be extrapolated to the PL processes, but it becomes less convincing in that case. Consider $\sum u_i(t)$, where each sequence $u_i(t)$ develops without influencing the other sequences .($u_i(t')$ and $u_j(t'')$ are independent, whichever t' and t'', if $i \neq j$; but may be dependent, if $i = j$.)

The limit of the normed sum of the u_i is necessarily bivariate stable. Under appropriate conditions, it is PL. The independence of the evolutions of the u_i may be questioned, in this motivation of the PL bivariate law.

Another type of motivation of the use of the PL variables and sequences is based on the stability property, without reference to the limit property. It is known, that one of the main difficulties in income studies, is the definition of income itself. Different observers may use different definitions, and some mix several definitions, because they must use other people's data, which are not sufficiently well described. Under the circumstances, any theory must, first of all, be invariant under a certain set of "linguistic" transformations, which express its invariance relative to the observer himself. When it is shown that the PL laws are the only ones to possess such invariance properties, one will have provided a kind of weak, "phenomenological" theory of these laws.

These additional arguments may be applied along different dimensions. As an example, one may consider either the incomes from different sources, or incomes before and after tax, or incomes of husband alone, or of husband and wife, etc....

Such arguments cannot be conveniently developed more fully in a report such as the present one.

One objection against these arguments may be that, in fact, one observes that each particular individual's income comes overwhelmingly from a single source, which differs from individual to individual. This is in contradiction to another widespread feeling, that each contribution to a sum must be small, relative to this sum. However, this "feeling" is only an expression of a theorem, which is valid only when the sum is a gaussian. If the sum is not gaussian, it is highly probable that a very large part of this sum comes from the single largest contribution. Only, a priori, one does not know which contribution will be

the largest. There is therefore no contradiction with intuition.

An interesting case is when one considers only the sum of two components. One of them will turn out to contribute most of the total earnings.

Therefore, the PL distribution may also be characterized, among all the possible limit distributions, as being the one, which corresponds to a finite mean value of the variable, together with extremely unequal partition between the different elements of income.

2.4. Properties of PL sequences. Random walks with a special kind of boundary layer. Return to the bivariate PL distribution, and, to make the argument simpler, assume that the function G(z) is composed of a finite number of discrete jumps. If u(t') and u(t") are independent, G(z) has only two jumps, for z = 0, and for z = $\pi/2$. In the case of dependence, there are also other jumps for G(z).

Suppose that the distribution of $\bar{U}$ = (u',u") is symmetrical. Then, $\bar{U}$ may be decomposed into two parts, one of which is a sequence of independent variables, and the other contains no jump of G(z) for z = 0 or z = $\pi/2$. Consider more specially this second part of $\bar{U}$, and let u(t) be a Markovian sequence.

From what is already known about u' and u", these variables are each a sum of a finite number of contributions, and if they are both large, then, most probably, a high part of each variable will come from the highest contribution. However, after the subtraction of the independent sequence, the contributing vectors are not perpendicular to either axis of coordinates; therefore, most of both u' and u" comes from the same contributing vector.

Suppose now that u' is already known, and that u" is a Markovian variable, depending only upon u'. One knows that u"/u' = tg(z), where z is one of the angles, for which G(z) has a jump. The probabilities of these values of z are independent

of u'. Finally, whichever the value of u', (if only it is very large) one goes from log u' to log u" by addition of one of the terms log tg(z), each of which having a well-determined probability.

This shows that, when the variables u', u" in a Markovian sequence, jointly follow a bivariate PL distribution, then, for large values of u, log u performs a random walk. It may also be shown that the average change of log u, or of u, is negative.

The above derivation, although rather heuristic, may be made rigorous. It is seen that the range of values of u, to which it applies, increases when the slopes of the vectors (u'u"), corresponding to the jumps of G(z), are neither too close to 0, nor to $\pi/2$. That is, it is the more accurate, the greater the dependence between u' and u".

For small values of u, the random walk model applies no more. But one needs not to invent ad hoc boundary conditions, as in the a priori Champernowne approach. On the contrary, this model provides a new type of limit layer for a random walk, which can be studied for its own merits.

It will be noted that, as in the ad hoc model, the index α will depend upon the probabilities of the different jumps in the random walk, and not on the nature of the boundary. But the number u^c will depend upon the boundary.

If u(t) is now taken to be a Markov chain of memory greater than 1, the random walk will be more complicated, but the essential spirit of the theory remains.

Is it possible to consider other models of the income change than the Markovian? The one which has been studied in greatest detail is the model involving moving averages. That is, it is assumed that the income at instant t is the weighted mean of a set of independent stable income impulses w, received at all the preceeding instants of time:

$$u(t) = \int w\,(t-h)\; g(h)$$

This model is easiest to generalize to the case when time is no longer taken as discrete. It should be noted that the usual theory of linear filtering of stationary (second order) processes is entirely powerless in the present case.

2.5. The problem of the interpolation of income time series. So far, the random walk approach seemed to be rather reasonable; but it does not remain so, when the problem of the time interval is raised.

First, consider the case when the PL laws hold and the incomes over non-overlapping time intervals are independent. The income over T may be divided into the sum of incomes over N intervals T/N. Most probably, most of this income will come from the best of the N intervals, if the whole is large. The interpolation may be continued until income becomes a function of increments over continuous time. It is known then that the income over T is essentially made out of discontinuous increments, and that there is anyway no sense in a rate of income over very short time intervals. The income over T is likely to come from some most favorable single jump.

Let now the income function depend upon past incomes, and let time become continuous. For large incomes, one may be tempted to interpolate from the random walk approximation. If so, one would conclude that increments of log u, over successive T/2 time intervals, were of the same order of magnitude. This is very much in contradiction with the case of independence.

However, such an interpolation would be basically wrong. This is so, because the interpolation of a random walk involves the interpolation of the Markovian hypothesis. One assumes that, whatever the time interval T/N, the income over the succeeding T/N will depend only upon the preceeding one, and will differ little from it, on the average. Therefore, the

regularity of income earning was built into the system. In fact, the interpolation of the Markovian hypothesis is basically wrong. At best, one may assume that the income over the decreasing time interval T/N only depends upon the total income over the preceeding T. A weaker assumption: The income over T/N may depend upon the whole curve of income variation, over the preceeding N time interval T. If so, nothing insures the continuity of u(t) any more, and, in fact, one comes closer to what is the case when the incomes are independent over distinct time intervals.

CHAPTER III. ESTIMATION - THEORETICAL FOUNDATIONS OF THERMOSTATISTICS

(Summary)

The main point of the new approach to the foundations of thermodynamics is that a central role is played by the bounds, which mathematical statistics imposes on certain processes of measurement, considered as estimation problem. In this approach, the concept of information, which has been on the basis of many recent attempts to better understand the foundations of thermodynamics, has been imbedded into a very much more detailed description of measurement. As a consequence, the foundations of thermodynamics are not only related to information theory, in Shannon's sense, (that is, to a method of designing signals) but are also related to detection theory (that is, to a method of handling signals which have already been designed.)

The main conceptual problem of thermodynamics arises, as is well known, from an uncompatibility of structure between the irreversibility of classical phenomenological thermodynamics, and the reversibility of any kinetic model one could ever think of. The problem remains in quantum mechanics. For example, all that Gibbs ever hoped to find, in the mechanics

of large assemblies of systems, were only "analogs" to thermodynamics. For that, one needs to add some hypotheses, which are fairly arbitrary, and are even today the object of discussion. In fact, one often cannot avoid the reference to some observer. However, once randomness has been introduced, it is preserved and its evolution in time can be followed up with little new conceptual difficulty, although with some great technical difficulties, in some cases.

The dynamic theory is therefore insufficient, if one does not add some probabilistic hypotheses. Is it then necessary, if such hypotheses are added? Could it not be possible to somehow short-circuit even the reference to atoms? Anyway, is there not a part of thermodynamics, which can be deduced from the only fact that one has somewhere postulated the existence of some probability distribution. (Note that if atoms are short circuited, one cannot speak of configurations, and apply the microcanonical equirepartion principle.)

The above problem has been set up by Szilard, in 1925, (at least implicitly) and he has shown that it is indeed possible to construct a theory, which is both statistical and phenomenological. Szilard's paper is often quoted, but never analyzed, and was perhaps ill understood. Similar remarks were made by G.N. Lewis in 1931. This author's independent investigations lead to a seemingly improved version of Szilard's results, and can be pushed further.

Case of one-parameter systems. Let all the properties of a system depend upon a single parameter $b = 1/T$, the dependence being stochastic. That is, for every "observable" U, one only knows a probability distribution $p(u/b)$. In that case, b itself cannot be considered as an "observable". It can only be "inferred", "estimated", from such observables as U. In other terms, there exists a certain type of physical quantities, which can be referred to as "estimables". It has been stressed by

Neyman, that estimation always includes a part of arbitrariness which decreases, as the size of "sample" increases.

Let U be additive. That is, the U of the sum of two systems is the sum of the U of each system alone. This is a hypothesis of stochastic independence. It is certainly true, if the systems are very large, the interaction U (energy) being negligible. Let one now look for reasonable ways of expressing the fact that b is a "macroscopic intensive variable of state". One may choose the following criteria: i) b is also the variable of state of any subsystem of the system considered; ii) the knowledge of b is in no way improved by the knowledge of the repartition of u among the subsystems of the system considered. That is, whichever the criterion of estimation that one may choose, one will find the same value for b, or the same confidence region, etc..., whether one knows the u of the whole, or the set of u's of all the parts of the system.

This criterion is made mathematically accurate by postulating:

"That u is a sufficient statistic for the estimation of b."

Entropy. The general theory was developed from these foundations, and it was shown that the canonical formula followed from it:

$$p(u/b) = s(u) \exp(-bu)\ g^{-1}(b)$$

In recent months, the problem of entropy was tackled. Entropy was taken to be the statistical counterpart of the expression $\int dQ/T$ of non-statistical thermodynamics.

First, canonical systems were considered. That is, T was a well-defined function of time, the system being put in thermal contact with a sequence of reservoirs, having the temperatures T (time). Then, dQ and dQ/T is a random variable, and $\int dQ/T$ may depend upon chance, and upon the whole path of the system, as parameter. However, it was shown that, with probability 1, the influence of the path followed is zero, and

that $dQ/T = -\log p e^{-bu} g^{-1}$

This is a method of introducing the "state" or configuration" of the system into the present theory. Besides, one finds that the random entropy, that is, the unaveraged -log p (state), does follow the second law of thermodynamics, despite statements to the contrary in many sources, for example in Khinchin.

Second, microcanonical systems were considered. That is, well defined quantities of heat were added to the system, T being at each instant of time given by some well-defined, but arbitrary, method of estimation. Then, for each definition of dQ, there corresponds a single well-defined definition of temperature, for which the second law holds without lower scale limit. If the system becomes large, all the estimation procedures converge, if they are "reasonable", and there is no difference between definitions anymore. All the usual physical systems are "very large", from this viewpoint.

Divisibility. Another problem being studied in greater detail is the problem of the divisibility of physical systems, into independent components.