The Journal of Political Economy: 71, 1963, 421-440. **E3**

New methods in statistical economics

• ***Chapter foreword***. An interesting relationship between the methods in this chapter and renormalization as understood by physicists is described in the *Annotation for the physicists* that follows this text. •

✦ **Abstract.** This is an informal presentation of several new mathematical approaches to the study of speculative markets and of other economic phenomena. My principal thesis is that to achieve a workable description of price changes, of the distribution of income, firm sizes, etc., it is necessary to use random variables that have an infinite population variance.

This work should lead to a revival of interest in Pareto's law for the distribution of personal income. The scaling distribution related to this law should dominate economics. ✦

AMONG TODAY'S STATISTICIANS AND ECONOMISTS, Pareto's law for the distribution of income is well-known, but is thoroughly neglected for at least two reasons. It fails to represent the middle range of incomes, and lacks theoretical justification within the context of elementary probability theory. I believe, however, that Pareto's remarkable empirical finding deserves a systematic reexamination, in light of the new methods that I attempt to introduce into statistical economics.

I. INTRODUCTION

Pareto claimed that there exist two constants, a prefactor C and an exponent $\alpha > 0$, such that for large u, the relative number of individuals with an income exceeding u can be written in the form $P(u) \sim Cu^{-\alpha}$.

That is, when the logarithm of the number of incomes greater than u is plotted as a function of the logarithm of u, one obtains for large u a straight line with slope equal to $-\alpha$. Later, the same relation was found to apply to the tails of the distributions of firm and city sizes. In fact, the search for new instances of straight log-log plots has been very popular and quite successful, among others, in Zipf 1941, 1949.

This book reserves the term "law of Pareto" to instances that involve the empirical distribution of personal income. The tail distribution $P(u) \sim Cu^{-\alpha}$ is denoted by the neutral term, *scaling distribution*, that is useable in many sciences and was not available when the paper reproduced in this chapter was published for the first time. The quantity α will be called *scaling exponent*.

Notwithstanding the abundant favorable evidence, Zipf's claims met strong objections from statisticians and economists. Those objections were so strong as to blind the critics to the evidence. In sharp contrast, I propose to show that *the scaling distribution literally cries* for our attention under many circumstances. Those circumstances include (1) taking seriously the simplified models based on maximization or on linear aggregation (2) taking a cautious view of the origin of the economic data or (3) believing that the physical distribution of various scarce mineral resources and of rainfall is important in economics.

In addition, I shall show that, when the "spontaneous activity" of a system is ruled by a scaling rather than a Gaussian process, the causally structural features of the system are more likely to be obscured by noise. They may even be completely "drowned out." This so because scaling noise generates a variety of "patterns;" everyone agrees on their form, but they have no predictive value. Thus, in the presence of a scaling "spontaneous activity, validating a causal relation must assume an unexpectedly heavy burden of proof and must acquire many new and quite perturbing features.

We shall see that the most important feature of the scaling distribution is the length of its tail, not its extreme skewness. In fact, I shall introduce a variant of the scaling distribution, which is two-tailed, and may even be symmetric. Hence, extreme skewness can be viewed as a secondary feature one must expect in variables that have one long tail and are constrained to be positive.

Much of the mathematics that I use as tool have long been available, but viewed as esoteric and of no possible use in the sciences. Nor is this paper primarily an account of empirical findings, even though I was the first to establish some important properties of temporal changes of speculative prices. What I do hope is that the methods to be proposed will constitute workable "keys" to further developments along a long-mired frontier of economics. Their value should depend on (1) the length and number of successful chains of reasoning that they have made possible; (2) the number of seemingly reasonable questions that they may show to be actually "ill-set" and hence without answer; and last, but of course not least, (3) the practical importance of the areas in which all these developments take place.

This paper will not attempt to treat any point exhaustively nor to specify all the conditions of validity of my assertions; the details appear in the publications referenced. Many readers may prefer to read Section VI before Sections II-IV. Section IX examines Frederick Macauley's important and influential critique of Pareto's law.

II. INVARIANCES; "METHOD OF INVARIANT DISTRIBUTIONS"

The approach I use to study the scaling distribution arose from physics. It occurred to me that, before attempting to explain an empirical regularity, it would be a good idea to make sure that this empirical identity is "robust" enough to be actually observed. In other words, one must first examine carefully the conditions under which empirical observation is actually practiced. The scholar observes in order to describe but the entrepreneur observes in order to act. Both know that most economic quantities can hardly ever be observed directly and are usually altered by manipulations. In most practical problems, very little can be done about this difficulty, and one must be content with whatever approximation of the desired data is available. But the analytical formulas that express economic relationships cannot generally be expected to remain unaffected when the data are distorted by the transformations to which we shall turn momentarily. As a result, a relationship will be discovered more rapidly, and established with greater precision, if it "happens" to be invariant with respect to certain observational transformations. A relationship that is noninvariant will be discovered later and remain less firmly established. Three transformations are fundamental to varying extents.

Linear aggregation, or simple addition of various quantities in their common natural scale. The distributions of aggregate incomes are better known than the distributions of each kind of income taken separately. Long-term changes in most economic quantities are known with greater precision than the more interesting medium-term changes. Moreover, the meaning of "medium term" changes from series to series; a distribution that is not invariant under aggregation would be apparent in some series but not in others and, therefore, could not be firmly established. Aggregation also occurs in the context of firm sizes, in particular when "old" firms merge within a "new" one. The most universal type of aggregation occurs in linear models that add the (weighted) contributions of several "causes" or embody more generally linear relationships among variables or between the current and the past values of a single variable (autoregressive schemes). The preference for linear models is of course based on the unfortunate but unquestionable fact that mathematics offers few workable nonlinear tools to the scientist.

There is actually nothing new in my emphasis on invariance under aggregations. It is indeed well known that the sum of two independent Gaussian variables is itself Gaussian, which helps use Gaussian "error terms" in linear models. However, the common belief that only the Gaussian is invariant under aggregation is correct *only* if random variables with infinite population moments are excluded, which I shall *not* do (see Section V). Moreover, the Gaussian distribution is *not* invariant under our next two observational transformations.

One may aggregate a small or a very large number of quantities. Whenever possible, "very large" is approximated by "infinite" so that aggregation is intimately related to the central limit theorems that describe the limits of weighted sums of random variables.

Weighted mixture. In a weighted lottery a preliminary chance drawing selects one of several final drawings in which the gambler acquires the right to participate. This provides a model for other actually observed variables. For example, if one does not know the precise origin of a given set of income data, one may view it as picked at random among a number of possible basic distributions; the distribution of observed incomes would then be a mixture of the basic distributions. Similarly, price data often refer to grades of a commodity that are not precisely known, and hence can be assumed to be randomly determined. Finally, the very notion of a firm is to some extent indeterminate, as one can see in the case of subsidiaries that are almost wholly owned but legally distinct. Available data often refer to "firms" that actually vary in size between individual establishments and holding companies. Such a mixture may be represented by random weighting. In many cases, one deals with a combination of the above operations. For example, after a wave of mergers hits an industry, the distribution of "new" firms may be viewed as a *mixture* of (a) the distribution of companies *not* involved in a merger, (b) the distribution of companies that are the *sum* of two old firms, and perhaps even (c) the *sum* of more than two firms.

Maximizing choice, the selection of the largest or smallest quantity in a set. It may be the case that all we know about a set of quantities is the size of the one chosen by a profit maximizer. Similarly, if one uses historical data, one must often expect to find that the fully reported events are the exceptional ones, such as droughts, floods or famines (and the names of the "bad kings" who reigned in those times) and "good times" (and the names of the "good kings"). Worse, many data are a mixture of full reported data and of data limited to the extreme cases.

Although the above transformations are not the only ones of interest, they are so important that it is important to characterize the distributions that they leave unchanged. It so happens that *invariance-up-to-scale holds asymptotically for all three transformations, as long as the parts themselves are asymptotically scaling*. In the case of infinite aggregation, invariance only holds if the scaling exponent α is less than two. To the contrary (with some qualifications), *invariance does not hold – even asymptotically – in any other case*.

Hence, anyone who believes in the importance of those transformations will attach a special importance to scaling phenomena, at least from a purely pragmatic viewpoint.

This proposition also affects the proper presentation of empirical results. For example, to be precise in the statement of scientific distributions, it is *not* sufficient to say that the distribution of income is scaling; one must list the excluded alternatives. A statistician will want to say that "it is true that incomes (or firm sizes) follow the scaling distribution; it is not true that incomes follow either Gaussian, Poisson, negative binomial or log-normal distributions" But my work suggests that one must rather say: "It is true that incomes (or firm sizes) follow the scaling distribution; it is not true that the distributions of income are very sensitive to the methods of reporting and of observation."

III. INVARIANCE PROPERTIES OF THE SCALING DISTRIBUTION

Of course, the invariance of the asymptotic scaling distribution holds only under additional assumptions; the problem will surely not be exhausted by the present approach. Consider N independent random variables, $U_n (1 \leq n \leq N)$ that follow the weak (asymptotic) form of the scaling distribution with *the same exponent* α. This means that

$$\Pr\{U_n > u\} \sim C_n u^{-\alpha} \text{ if } u \text{ is large.}$$

The behavior of $\Pr\{U_n < -u\}$ for large u will be examined in Section VII.

Let me begin with mathematical statements that imply that the scaling behavior of U_n is *sufficient* for the three types of asymptotic invariance-up-to-scale. Short proofs will be given in parentheses, and longer ones in the Appendix. The symbol Σ will always refer to the addition of the terms relative to the N possible values of the index n.

Weighted Mixture. Suppose that the random variable U_W is a weighted mixture of the U_n, and denote by p_n the probability that U_W is identical to U_n. One can show that this U_w is also asymptotically scaling and that its scale parameter is $C_W = \Sigma p_n C_n$, which is simply the *weighted average* of the separate scale coefficients C_n. (*Proof.* It is easy to see that

$$\Pr\{U_W > u\} = \sum p_n \Pr\{U_n > u\} \sim \sum C_n p_n u_n^{-\alpha} = C_w u^{-\alpha}.)$$

Maximizing choice. *Ex-post*, when the values U_n of all the variables U_n are known, let U_M be the largest. One can show that this U_M is also asymptotically scaling, with the scale parameters $C_M = \Sigma C_n$, the *sum* of the separate scale coefficients C_n. (*Proof.* Clearly, in order that $U_M \leq u$, it is both necessary and sufficient that $U_n \leq u$ is valid *for every* n. Hence, Π denoting the product of the terms relative to the N possible values of the index n, we have

$$\Pr\{U_M < u\} = \Pi \Pr\{U_n \leq u\}.$$

It follows that

$$\Pr\{U_M > u\} = 1 - \Pr\{U_M \leq u\} \sim 1 - \Pi(1 - C_n u^{-\alpha}) \sim \sum C_n u^{-\alpha} = C_M u^{-\alpha}.)$$

Aggregation. Let U_A be the sum of the random variables U_n. One can show that it is also asymptotically scaling, with a scale parameter that is again the *sum* of the separate weights C_n. Thus, at least asymptotically for $u \rightarrow \infty$, the sum of the U_n behaves exactly like the largest U_n (see M 1960i{E10} for further details). Mixture combined with aggregation is an operation that occurs in the theory of random mergers of industrial firms

(M 1963o). One can show that it also leaves the scaling distribution invariant-up-to-scale.

The converses of the above statements are true only in the first approximation; for the invariance-up-to-scale to hold, the distributions of the U_n need not follow the scaling distribution exactly; but they must be so close to it as to be scaling for many practical purposes.

Strictly invariant distributions that also enter as limits. To introduce two distributions due to Fréchet and Lévy, respectively, and relate them to scaling, let us imitate (with a different interpretation) a principle of invariance that is typical of physics: We shall require that the random variable U_n be strictly invariant up to scale with respect to one of our three transformations.

Let N random variables U_n follow – up to changes of scale – the same distribution as the variable U, so that U_n can be written as a_nU, where $a_n > 0$. I shall require that U_W (respectively, U_M or U_A) also follow – up the changes of scale – the same distribution as U. This allows one to write U_W (U_M or U_A) in the form a_WU (a_MU or a_AU), where a_w, a_m and a_A are positive functions of the numbers a_n.

As shown in the Appendix, it turns out that the conditions of invariance lead to somewhat similar equations in all three cases; ultimately, one obtains the following results:

Maximization. The invariant distributions must be of the form $F_M(u) = \exp(-u^{-\alpha})$ (Fréchet 1927, Gumbel 1958). These distributions are clearly scaling for large u and correspondingly small $u^{-\alpha}$, since in that range F_M can be approximated by $1 - Cu^{-\alpha}$. They also "happen" to have the remarkable property of being the limit distributions of expressions of the form $N^{-1/\alpha} \max U_n$, where the U_n are asymptotically scaling. There are no other distributions that can be obtained simply by multiplying the mass U_n by an appropriate factor and by having N tend to infinity. But allowing the origin of U to change as $N \rightarrow \infty$, yields the "Fisher-Tippett distribution," which is *not* scaling and not invariant under the other two transformations.

Mixing. In this case, the invariant distributions are $F_W(u) = 1 - Cu^{-\alpha}$, which is the analytical form of the scaling distribution extended down to $u = 0$. This solution corresponds to an infinite total probability, implying that, strictly speaking, it is unacceptable. However, it must not be rejected immediately because in many cases U is further restricted by some relation of the form $0 < a \leq u \leq b$, leading to a perfectly acceptable conditional probability distribution.

Aggregation. Finally, aggregation leads to random variables that are the "positive" members of the family of "L-stable distributions," other members of which will be encountered =later (Lévy 1925, Gnedenko & Kolmogorov 1954). These distributions depend on several parameters, the principal of which is again denoted by α and must satisfy $0 < \alpha \leq 2$. The density $dF_A(u)$ has a closed analytic form in a few cases. The limit case for $\alpha = 2$, is the Gaussian distribution (which, however, is not itself scaling). The density of the positive L-stable distribution is also known in the case $\alpha = 1/2$, which plays a central role in the study of the return to equilibrium in coin tossing. In other cases, no closed analytic expression is known for the stable distribution $F_A(u)$. But Lévy showed that they asymptotically follow the scaling distribution with exponent α, except in the limit case $\alpha = 2$ (for α just below 2, their convergence to their scaling limit is slow).

The L-stable variables yielded by the present argument can take negative values if $1 \leq \alpha \leq 2$, as is readily seen in the Gaussian case. But there is a very small probability that they take *large* negative values. I have shown how this can be handled in practice by suitably displacing the origin.

L-stable distributions have another important property: they are the only possible non-Gaussian limits of linearly weighted sums of random variables. Hence, even though they cannot begin to compare with the Gaussian from the viewpoint of ease of mathematical manipulations, they both share the fundamental properties of that distribution from the viewpoint of linear operations. The corresponding forms of the non-classical central limit theorem show that the sum of many additive contributions need *not* be Gaussian; if one wishes to explain by linear addition a phenomenon that is ruled by a skew distribution, it is *not* necessary to assume that the addition in question is performed in the scale of U itself. This also shows that the log-normal distribution is *not* the only skew distribution that can be explained by addition arguments, thus removing the principal asset of that distribution (which is known in most cases to underestimate grossly the largest values that can be taken by the variable of interest).

One can see that the probability densities of the three invariant families differ throughout most of the range of u. However, if $0 < \alpha < 2$, their asymptotic behaviors coincide. Hence, the scaling distribution is also asymptotically invariant with respect to applications of an arbitrary succession of the basic transformations. When α is close to 2, the practical application of this property requires additional qualifying statements.

It should be noted that Fréchet's and Lévy's distributions attract substantial attention from mathematicians. However, the scaling maximum distributions have few generally known applications and the scaling sum distributions (L-stable distributions) have practically none.

It is true that a celebrated treatise on stable distributions, Gnedenko & Kolmogorov 1954, alludes to forthcoming publications specifically concerned with applications of L-stability. However, when I discussed this allusion with Professor Kolmogorov in 1958 (ten years after the original Russian edition,) I found that these papers had not materialized after all – for lack of applications! Basically, the only fairly well-known practical instance of a stable distribution is the distribution due to Holtsmark (but often rediscovered,) which rules the Newtonian attraction between randomly distributed stars (see Section 2.8 of M 1960i{E10}). Thus, Gnedenko & Kolmogorov 1954 did not pre-empt my plea that stable distributions should be counted among the most "common" probability distributions.

IV. SIGNIFICANCE OF THE EVIDENCE PROVIDED BY DOUBLY LOGARITHMIC GRAPHS

Limitations on the value of *a* lead to another quite different aspect of the general problem of observation. It concerns the practical significance of statements having only an asymptotic validity. Indeed, to verify empirically the scaling distribution, the usual first step is to draw a doubly logarithmic graph: a plot of $\log_{10}[1-F(u)]$ as a function of $\log_{10}u$. One should find that this graph is a straight line with the slope $-\alpha$, or at least that it rapidly becomes straight as u increases. But, look closer at the sampling point of the largest u. Except for the distribution of incomes, one seldom has samples over 1,000 or 2,000 items; therefore, one seldom knows the value of u that is exceeded with the frequency $1-F(u)=1,000^{-1}$ or $2{,}000^{-1}$. That is, the "height" of the sampling doubly logarithmic graph will seldom exceed *three* units of the decimal logarithm of $1-F$. The "width" of this graph will be at best equal to $3/\alpha$ units of the decimal logarithm of u. However, if one wants to estimate reliably the value of the slope *a*, it is necessary that the width of the graph be close to one unit. In conclusion, one cannot trust any data that suggest that α is larger than 3. Observe that the resulting practical range of α's is wider than in the case of stable distributions.

Looking at the same question from another angle, take doubly logarithmic paper and plot the following distributions: Gaussian, lognormal, negative binomial and exponential. Because all these distributions are very "short tailed," the slope of the graph will become asymptotically infinite. However, in the region of probabilities equal to one-thousandth, the dispersion of sample data is likely to generate – on doubly logarithmic coordinates – the appearance of a straight line having a high but finite slope. In the words of Macaulay 1922 (see Section IX): "The approximate linearity of the tail of a frequency distribution charted on a double logarithmic scale signifies relatively little, because it is such a common characteristic of frequency distributions of many and various types." However, linearity with a low slope signifies a great deal indeed. Figure 1 further illustrates this difference between different values of α.

There is another way to describe curve-fitting using special paper. One may say that the maximum distance between the sample curve and some reference curve – preferably a straight line – defines a kind of "distance" between two alternative probability distributions. Any special paper, whether it be log-normal or scaling, should be used only in ranges where the distances that it defines are sensitive to the differences that matter to the particular problem. Hence, the most conservative approach is often to consider several hypotheses, that is, to use several kinds of paper.

In summary, if one considers mixtures, maximizations and practical measurement, the range of values of α is reduced to the interval from 0 to 3. If one also takes aggregation into account, α must fall between 0 and 2 (actually, the range of "apparent" α's is somewhat wider).

V. FINITE SAMPLE BEHAVIOR OF RANDOM VARIABLES WITH INFINITE POPULATION MOMENTS

When α is not small (in a sense we shall describe shortly), a scaling distribution is extraordinarily long-tailed, as measured by Gaussian standards. In particular, if $\alpha < 2$, the population second moment is infinite. It should be stressed, however, that the concept of infinite variance is in no way "improper."

It is of course true that, since observed variables are finite, the sample moments of all orders are themselves finite for finite sample sizes; but this does not exclude the possibility that they tend to infinity with increasing sample size. It may also be true that the asymptotic behavior of the samples is practically irrelevant because the sizes of all empirical samples are by nature finite. For example, one may argue that the history of cotton prices is mostly a set of data from 1816 to 1958, speculation on cotton having been very much decreased by the 1958 acts of the United States Congress. Similarly, when one studies the sizes of United States cities, the statistical populations have a bounded sample size. Even for continuing series, one may well argue for "après moi, le déluge" and neglect any time horizon longer than a man's life. Hence, the behavior of the moments for infinite sample sizes may seem unimportant. But it actually implies that the only meaningful consequences of infinite population moments are those relative to the sample moments of increasing *subsets* of our various bounded universes.

In Figure 2, the predictions of the mathematical theory are illustrated by computer simulations. Distinct samples of scaling random variables with $\alpha=1$ were obtained by inverting samples of random variables dis-

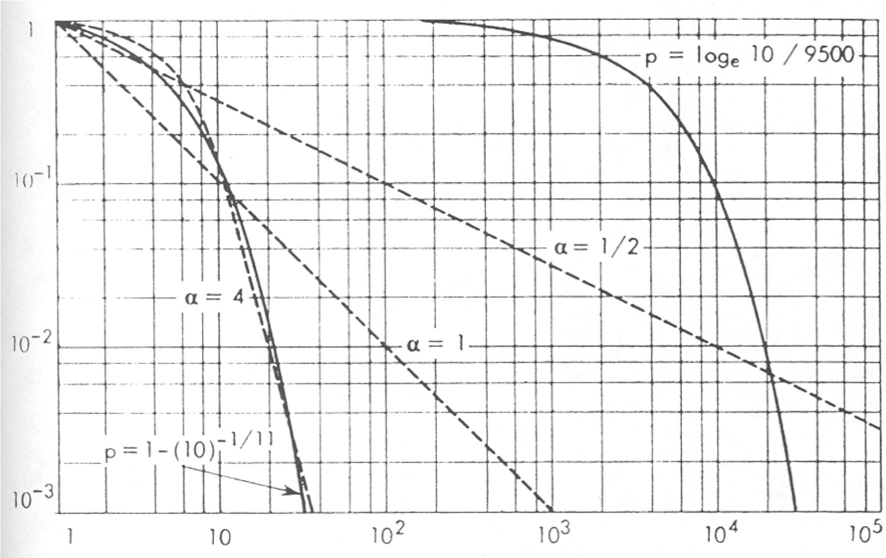

FIGURE C3-1. Five doubly logarithmic plots: (A) Two exponential distributions (*very curved solid lines*) with very different expectations. (B) Two distributions which are uniformly scaling from $u=1$ and have, respectively, the exponents $\alpha=1/2$ and $\alpha=1$. (C) One asymptotically scaling distribution, with the exponent $\alpha=4$, a large value. The relations between these graphs demonstrate graphically that distributions similar to (C) can readily be confused with the exponential, but small values of the α exponent are reliable.

tributed uniformly over the interval [0, 1]. Plots of the variation of the first and second moments are then created. The sample first moments illustrate what happens when the population moment is given by a barely divergent integral; the sample second moments illustrate what happens when the population moment is given by a rapidly divergent integral. The sample moments do not converge, and – even more impressive – their growth is erratic and very sample-dependent.

Let us now return to experimental data. In some cases, the sample second moment is observed to "stabilize" rapidly around the final value corresponding to the total set. If so, it is unquestionably useful to take this final value as an estimate of the population second moment of a conjectural infinite population from which the sample could have been drawn. But Figure 3 shows that the sample second moments corresponding to increasing subsets may continue to vary widely even when the sample size approaches the maximum imposed by the subject matter. From the

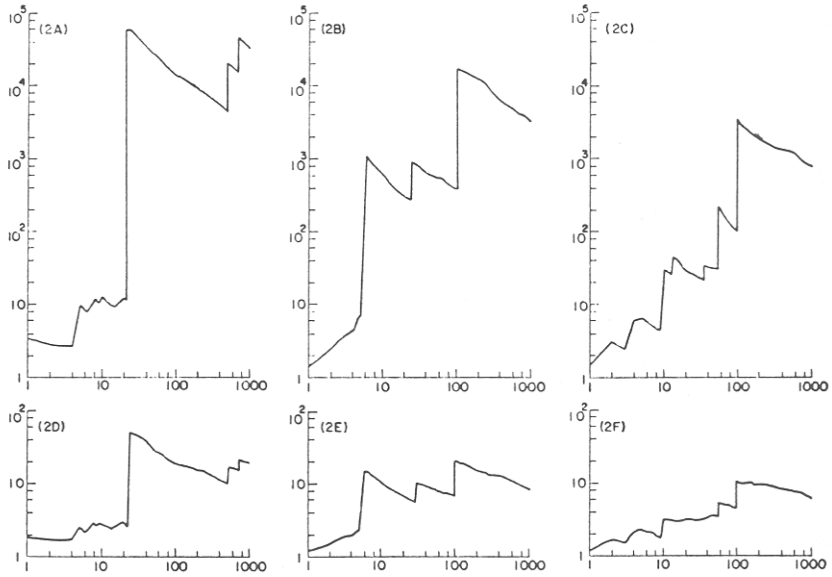

FIGURE C3-2. Monte Carlo runs of the sequential first moment (*lower graphs*) and the sequential second moment (*upper graphs*) of three independent samples from a scaling population of exponent $\alpha = 1$. The term "sequential moment" means that, in each run, the moment is computed for every sample size from 1 to 1,000. This figure suggests the degree to which the sample moments of scaling variables can be erratic and sample-dependent.

viewpoint of sampling, this expresses that even the largest available sample is too small for reliable estimation of the population second moment. In other words, a wide range of values of the population second moment are equally compatible with the data. Now, let us suppose that – as in Figure 3 – the appearance of the sample data recalls Figure 2. Then, the reasonable range of values for the population moment will frequently include the value "infinity," implying that facts can be equally well described by assuming that the "actual" moment is finite but extremely large or by assuming that it is infinite.

To support the alternative that I prefer, let me point out that a realistic scientific model must not depend too critically on quantities that are difficult to measure. The finite-moment model is unfortunately very sensitive to the value of the population second moment, and there are many other ways in which the first assumption, which of course is the more reasonable *a priori*, is also by far more cumbersome analytically. The second

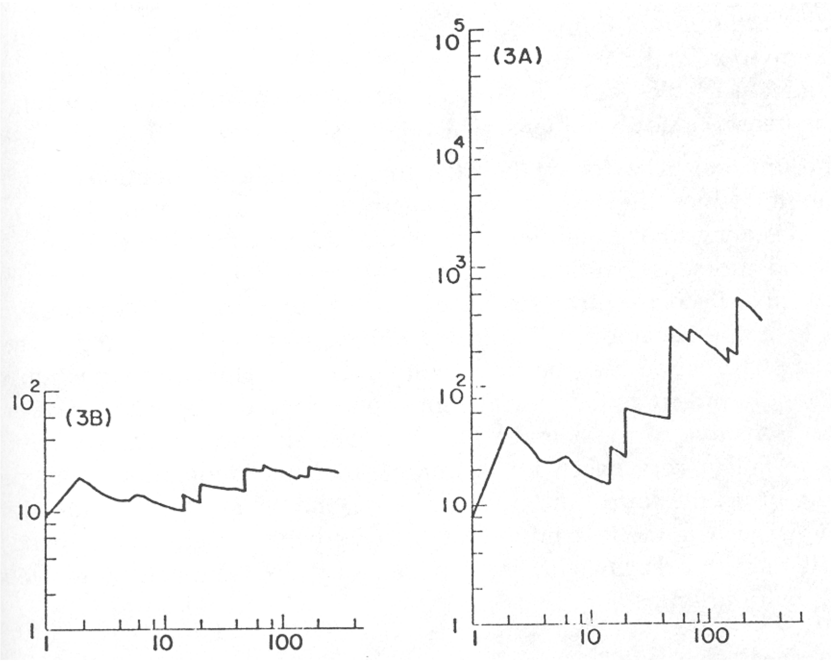

FIGURE C3-3. Sequential first moment (*left*) and the sequential second moment (*right*) of the numbers of inhabitants in United States cities with over 50,000 inhabitants. The cities have been ordered alphabetically. As city sizes have a scaling exponent of about $\alpha = 1.1$, the sample first moment tends – very slowly – to a limit, while the second moment increases less rapidly than in the simulations reported in Figure 2.

assumption, on the contrary, leads to simple analytical developments, and the rapidity of growth of the sample second moment can be modulated to lead to absurd results only if one applied it to "infinite" samples, that is, if one raised problems devoid of concrete meaning.

In other words, there is nothing absurd in assuming, as I am constantly led to do, that intrinsic bounded variables are drawn at random from infinite populations of unbounded variables having an infinite second moment. But all these infinities are a relative matter, entirely dependent on the statistician's span of interest. As the maximum useful sample size increases, the range of the estimates of the second moment will steadily narrow. Hence, beyond a certain limit, the second moments of some variables may be considered as finite. Conversely, there are variables for which the second moment must be considered finite only if the useful sample size is smaller than some limit.

Actually, this use of infinity is common in statistics, insofar as it concerns the function max $\{u_1, u_2, ..., u_N\}$ of the observations. From this viewpoint, even the use of infinite spans would seem improper. However, it is well known in statistics that little work could be done without using unbounded variables. One even uses the Gaussian distribution to represent the height of adult humans, which is surely positive!

The unusual behavior of the moments of scaling distributions can be used to introduce the least precise interpretation of the validity of the notion of scaling. For example, suppose that the first moment is finite, but the second moment is infinite. Then, as u tends to infinity, the function $1 - F(u)$ must decrease more slowly than $1/u^2$ but more rapidly than $1/u$. In this case, the behavior of $F(u)$ in the tails is very important, and a very useful approximation may be $Cu^{-\alpha}$, with $1 < \alpha < 2$. This approximation is completely harmless as long as one limits oneself to consequences that are not very sensitive to the actual value of a. The situation is very different when the tail is very short, for example, when the population moments are finite up to the fourth order. In that case, the behavior of the function $F(u)$ for large u is far less important than its behavior elsewhere; hence, one will risk little harm with interpolations by the Gaussian or the lognormal distribution.

VI. DIFFICULTIES CONCERNING STATISTICAL INFERENCE AND CONFIRMATION OF SCIENTIFIC DISTRIBUTIONS, WHEN THE ERRORS (THAT IS, THE "BACKGROUND NOISE") ARE SCALING

It is well known that second moments are heavily used in statistical measures of dispersion, or "standard deviation," and in "least-squares" and "spectral" methods. Hence, whenever the considerations of Section V are required to explain the erratic behavior of sample second moments, a substantial portion of the usual methods of statistics should be expected to fail. Examples of such failures have, of course, often been observed empirically and may have contributed to the disrepute in which many writers hold the scaling distribution; but it is clearly unfair to blame a formal expression for the complications made inevitable by the data that it represents. If $2 < \alpha < 3$, second moments exist, but concepts based on third and fourth moments, – for example Pearson's measures of skewness and kurtosis – are meaningless.

I am certain that for practical purposes some of those difficulties eventually will be solved. However, as of today, they are so severe that we must reexamine the meaning of the popular but vague concept of "a structure." It is indeed a truism, especially in fields where actual experimentation is impossible, that one must carefully distinguish between patterns that can only be used for "historical" description of his records and those that are also useful for forecasting some aspect of the future. A useful vocabulary considers the search for distributions a kind of extraction and identification of a "signal" in the presence of "noise." In particular, as we have seen, modern inference theory teaches us always to list both the accepted and the rejected possibilities. The scientist's major problem is frequently to determine whether a conjectured "relation" is statistically significant with respect to what may be generally called "spontaneous activity," which is the resultant of all the influences that one cannot or does not want to control in the problem at hand and which is conveniently described with the help of various stochastic models.

It is not enough, however, that all members of a cultural group agree on the patterns that they read into a historical record. Indeed, although there is unanimity in the interpretation of *certain* Rorschach inkblots, they have no significance from the viewpoint of science as a system of *predictions*. Broadly speaking, a pattern is scientifically significant when it is felt to have a chance of being repeated, meaning that, in some sense, its "likelihood" of having occurred by chance is very small. Unfortunately, the tools of statistics have been mostly designed to deal with Gaussian alternatives and, when the chance alternative is scaling, they are not *at all* conservative or "robust" enough. One will often be able to circumvent this difficulty, but not always. In fields where the background noise is scaling, the burden of proof is closer to that of history and autobiography than of physics.

The same thought can be presented in more optimistic terms by saying that, if "mere chance" can so readily be confused with a causal structure, the effect of chance is itself entitled to be called a structure. The word "noise" may perhaps be reserved for the Gaussian error terms, or its binomial or Poisson kinds, which are seldom respected as sources of anything that looks interesting.

The situation is worse in models known to be very structured (for example, to be autoregressive) with scaling noise. Compared to the case of Gaussian noise, one should expect the data to be *much* more influenced by the noise and *much* less influenced by the structure.

The association between the scaling distribution and "interesting patterns" is nowhere more striking than in the game of tossing a fair coin, which Henry and Thomas have been playing since sometime in the early eighteenth century. When the coin falls on "heads," Henry wins a dollar (or perhaps rather a thaler); when the coin falls on "tails," Thomas wins. We disregard what happened to the game before we break in at time $t=0$, and we denote by T the time it takes for Henry and Thomas's fortunes to return to the state that they were in at the moment when we broke into the game. For large values t of T, one has the well-known relation: (Feller 1950, Vol. 1).

$$\begin{aligned}&\text{Probability } \{ \text{ that the fortunes return to their initial states} \\ &\qquad\qquad \text{after a time greater than } t\} = (\text{constant})\ t^{-1/2}.\end{aligned}$$

This relation involves the scaling distribution with exponent $\alpha = 1/2$.

However, gamblers are notorious for seeing an enormous amount of interesting detail in the past records of *accumulated* coin-tossing gains; far more than in the non-cumulative sequences. That is, gamblers are prepared to risk their fortunes on the proposition that these details are not due to mere chance. Several of my papers were based on the idea that very similar phenomena should be expected whenever the scaling distribution applies. If so, one could associate with those phenomena some stochastic models that dispense with any kind of built-in causal structure and yet generate sample curves in which both the unskilled and the skilled eye can distinguish the kind of detail that is usually associated with causal relations. In the case of Gaussian processes, such details would be so unlikely that they would surely be considered significant for forecasting; but, this is not true in the scaling case. From the viewpoint of prediction, those structures should be considered *perceptual illusions*: they are in the observer's current records and in his brain but not in the mechanism that has generated these records and that will generate the future events.

Bearing in mind the existence of such models, let us suppose that we have to infer a process from the data. A non-structured scaling universe accounts very well for many observations; as a result, it is extremely difficult, at best, to choose between it and an alternative model that postulates causal relations. It is very difficult to challenge someone's belief in the existence of "genuine" structures. But to communicate such a belief to others, with the standards of credibility that are current in physical science, requires *much* more than the statistical tests of significance that social scientists shrug off at the end of a discussion. Such a situation requires a drastic sharpening of the distinction between patterns that – however great the scholar's diligence – can serve only for historical purposes and those patterns that are useable for forecasting.

The question that I have in mind can be well illustrated by the problem of the significance of "cycles." Both the eye and sophisticated methods of Fourier analysis, suggest that almost any record of the past is a sum of periodic components. But the same is also true for a wide variety of artificial series generated by random processes with no built-in cyclic behavior. Furthermore, skilled cycle researchers seldom risk firm, short-term forecasts. Could we then ask two questions that paraphrase Keynes's comments on early econometric models, "How far are these curves meant to be no more than a piece of historical curve-fitting and description, and how far do they make inductive claims with reference to the future as well as the past?"

It may also be noted that, because of the invariance of the scaling distribution with respect to various transformations (see Section III), one cannot hope that a simple explanation will be provided by arguing that only the genuine structures will be apparent to all observers. The only criterion of trustworthiness is replicability in time.

In an important way, the models of scaling spontaneous activity differ from the standards of "operationalism" suggested by philosophers. Indeed, to explain by mere chance any given set of phenomena, it will be necessary to imbed them in a universe that also contains such a fantastic number of other possibilities that billions of years may be necessary to realize all of them. Hence, within our lifetime, any given configuration will occur at most once, and one could hardly define a probability on the basis of sample frequency. This conceptual difficulty is common knowledge among physicists, and it is to be regretted that the philosophical discussions of the foundations of probability seldom investigate this point. In a way, the physicists freely indulge in practices that for the historian are mortal sins: to rewrite history as it would have been if Cleopatra's nose had a different shape. My sins are even worse because their actual histories turn out to be very close to some kind of "norm," a property which my models certainly do not possess.

The foregoing argument is best illustrated by two separate reinterpretations of the coin-tossing record plotted in Figure 4. First, forgetting the origin of that figure, imagine that it is a geographical cross-section of a new part of the world in which all the regions below the bold horizontal lines are under water. Imagine also that this chart has just been brought home by an explorer; the problem is to decide whether it was due to cause or to chance. The naive defense will resort to the Highest Cause. Presenting our graph as fresh evidence that God created Heaven and Earth using a single template, it follows that such concepts as a "continent," an "ocean," an "island," an "archipelago" or a "lake" are precisely adapted to the shape of the Earth. However, a devil's advocate would argue that the Earth is a creation of blind chance and that the possibility of using such convenient terms as "continent" and "island" just reflects the fact that the areas above water happen often to be very short or very long and are rarely of average length.

The preceding example is not as fictitious as it may seem: the distribution of the sizes of actual islands happens to be scaling (M 1962n). Hence, our hypothetical debate emphasizes the two extreme viewpoints realistically, even though – the Earth having been presumably entirely explored – no actual prediction is involved in the choice between the interpretation of archipelagoes as "real" or as creations of the mind of the weary mariner.

Another example, also chosen for its lack of *direct* economic interpretation, is the problem of clusters of errors on telephone circuits. Suppose that a telephone line is used only to transmit either dots or dashes, which may be distorted in transmission to the point of being mistaken for each other. It is clear – again, according to the defender of a search for causes – that whenever an electrician touches the line, one should expect to observe a small cluster of such errors. Moreover, since a screwdriver touches the line many times during a single repair job, one should expect to see clusters of clusters of errors and even clusters of higher order.

Actual records of the instants when errors occurred do indeed exhibit such clusters in between long periods of flawless transmission. A good idea of the distribution of the errors is provided by yet another look at Figure 4. Consider the sequence of points where the graph crosses the

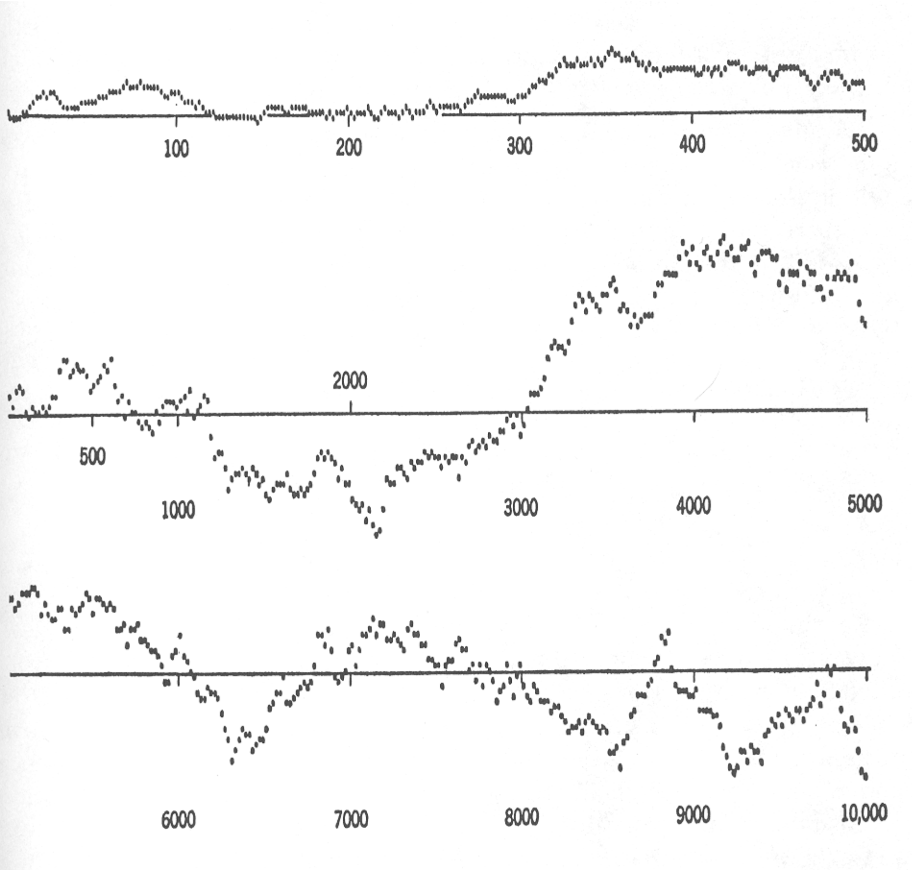

FIGURE C3-4. Record of Henry's winnings in a coin-tossing game, played with a fair coin. Zero-crossings seem to be strongly clustered, although intervals between crossings are obviously statistically independent. This figure is reproduced from Feller 1950 (Volume I).

To appreciate fully the extent of apparent clustering in this figure, note that the unit of time is 2 coin tosses on the first line, and 20 coin tosses on the second and third lines. Hence, the second and third lines lack detail and each apparent zero-crossing is an imperfect representation of a cluster or a cluster of clusters. For example, the details of the cluster centered around the 200th coin toss are clearly separated on line 1.

line that, in an earlier interpretation, had represented sea level. According to the searcher for causes, the precise study of such past records will improve the prediction of errors and will help minimize their effects. On the other hand, precisely because of the origin of Figure 4, those beautiful hierarchies of degrees of clustering can very well be due to a "mere chance" devoid of memory (see Berger & M 1963).

Similar devil's advocates can be heard in many contexts, and someone should take this role in relation to every important problem, without forgetting that the devil's advocate must always be on the side of the angels. An interesting example of a stable truce between structure and chance is provided by the study of language and of discourse, where the traditional kind of structure is represented by grammar and – as one should expect by now – the chance mechanism is akin to the scaling distribution (see Apostel, M & Morf 1957 and M 1961b).

VII. TWO-TAILED AND/OR MULTIDIMENSIONAL STABLE DISTRIBUTIONS

Until now, we have followed tradition by associating the scaling distribution with essentially positive random variables, the distribution of which has a single long tail, making its central portion necessarily quite skew. However, I have discovered important examples in economics of distributions having *two* scaling tails; the most striking example is that of relative changes in the prices of sensitive speculative commodities. The argument of invariance under maximization cannot extend to them. But invariance under mixture simply leads to the combination of the scaling distribution of positive u and the scaling distribution of negative u. Invariance under aggregation is satisfied by every random variable constructed by adding or subtracting two arbitrarily weighted "positive" stable variables of the kind studied earlier in this paper. In particular, these general stable variables can by symmetric; the Cauchy distribution provides a prime example. But their study depends very little on the actual degree of skewness. Hence, the asymmetry of the usual scaling variables is less crucial than the length of their single tail.

Another remarkable property of the stable distributions is that, like the Gaussian, they have intrinsic extensions to the multivariant case, other than the degenerate case of independent coordinates. Very few other distributions (if any) share this property. The reason for this is innately related to the role of stable distributions in linear models. It is indeed possible to characterize the multivariate stable distributions as being those for which the distribution of every linear combination of the coordinates is a scalar stable variable. This property is essential to the study of multidimensional economic quantities, as well as to the investigation of the dependence between successive values of a one-dimensional quantity, such as income (see M 1961e{E11}).

VIII. THE ROLE OF THE SCALING DISTRIBUTION IN ECONOMICS AND A LINK WITH THE PHYSICAL SCIENCES

The arguments of this paper show that there is a strong pragmatic reason to undertake the study of scaling economic distributions and time series. This category includes prices (M 1963b{E14}), firm sizes (M 1963o) and incomes (M 1960i{E10}, as amended in M 1963p, and also M 1967j{E15}, 1962g), hence making the study of scaling of fundamental importance in economic statistics. Similarly, the example of the distribution of city sizes stresses the importance of the scaling distribution in sociology (M 1965m). Finally, strong indications exist of its importance in psychology, but I shall not even attempt to outbid George Kingsley Zipf in listing all the scaling phenomena of which I am aware; their number seems to increase all the time.

However, it is impossible to postpone "explanation" forever. If indeed a grand economic system is only based on aggregation, choice and mixture, one can prove that for a system to be scaling, it *must* be triggered somewhere by essentially scaling "initial" conditions. That is, however useful the method of invariants may be, it is true that it somewhat begs the question and that the basic mystery of scaling cannot be solved by pushing around the point where such behavior is postulated. Indeed, if it were true, in accordance with "conventional wisdom," that physical phenomena are characterized by the distributions of Gauss and social phenomena are characterized by that of Pareto, we may eventually have to explain the latter using the "microscopic" economics models, such as the "principle" of random proportionate effect, which I prefer not to emphasize in my approach.

I claim, however, that this situation *need not* be the case. Quite to the contrary, the physical world is full of scaling phenomena that one can easily visualize as playing the role of the "triggers" that cause the economic system to be also scaling. For example (M 1962n), I have found that single-tailed scaling distributions, with trustworthy values for α, represent the statistical distributions of a variety of mineral resources, which are surely not influenced by the structure of society. This is the case with the areas of oil fields and the sums of their total past production and their currently estimated capacity). The same is true for the valuations of certain gold, uranium and diamond mines in South Africa. Similar findings hold for a host of similar data related to weather, which is barely influenced by man as yet. Some weather data, such as hail records, have a direct influence on important risk phenomena, namely, insurance against hail damage. Other weather data, such as total annual rainfall, obviously influence the sizes of crops and hence, by the distributions of supply and demand, influence the changes of agricultural prices.

If this paper proposed to contribute to "geo-statistics," it should, of course, examine the degree of generality of my claim. But, for the purpose of a study of economic time series, it will be quite sufficient to note that the trigger of a scaling grand economic system *can very well* be found in statistical features of the physical world. For example, natural resources and weather influence prices, which in turn influence incomes. Since the systems to which we refer are spatio-temporal, there is nothing disturbing in our association of economic *time* series with geological and geographical *spatial* distributions.

I shall not attempt to say anything about the actual triggering mechanism since I doubt that a unique link can be found between the social and the physical worlds. After all, quite divergent values of the scaling exponent α are encountered in both worlds so that the overall grand system cannot possibly be based only upon transformations by linear aggregation, choice and mixture.

I wish, finally, to point out that the scaling phenomena of physics have also turned out to include some phenomena with no direct relation with economics. For example, Section 3 mentioned that a three-dimensional stable distribution occurs in the theory of Newtonian attraction. Moreover, the distribution of the energies of the primary cosmic rays has long been known to follow a distribution that happens to be identical to that of Pareto with the exponent 1.8 (Fermi's study of this problem includes an unlikely but rather neat generation for the scaling distribution). The same result holds for meteorite energies, an important fact for ionospheric clatter telecommunications. Also, as discussed in Section VI, the intervals between successive errors of transmission on telephone circuits happen to be scaling with a very small exponent, the value of which depends on the physical properties of the circuit.

There are many reasons for believing that many scaling phenomena are related to "accumulative" processes similar to those encountered in coin-tossing.

IX. FREDERICK MACAULAY'S CRITICISM OF PARETO'S LAW

Having accumulated so many reasons to view the scaling distribution as extraordinarily important, I am continuously surprised by the attitude described in the first sentence of Section I. I eventually realized that it had deep roots not only in the apparent lack of theoretical motivation for that distribution but also in several seemingly "definitive" criticisms, such as that of Macaulay 1922.

Macaulay's essay is most impressive indeed and – even though I disagree with its conclusions – I strongly recommend it. It disposed of the claim that the α exponent in Pareto's law is the same in all countries and at all times and of the claim that the scaling distribution describes small incomes or the incomes of the lower paid professional categories. Macaulay is also very convincing concerning scaling distributions with a high exponent (see Section V).

I believe, however, that his strictures against "mere curve fitting" have been very harmful. His ideal of a proper mathematical description is so restrictive that he rejects the scaling distribution outright because the sample empirical curves do not "zigzag" around the simple scaling interpolate but rather cross it systematically a few times. This illustrates a basic difference between the care economists bring to statistics and the seeming carelessness of the physicists. For example, when the Boyle law was found to differ from the facts, the physicists simply invented the concept of a "perfect gas," that is, a body that follows Boyle's law *perfectly*. Naturally, perfect gas approximations are absurd in some problems but are adequate in many others, and they are so simple that one must consider them first. Similarly, scaling distribution approximations should not even be considered in problems relating to low incomes, but in other investigations they deserve to be the first to be considered.

Therefore I can summarize Macaulay's criticism of the scaling distribution by saying that it only endorses the asymptotic forms. In many cases, however, I believe that it is legitimate to consider more seriously certain "relatives" of the scaling distribution, such as the stable distributions.

APPENDIX: SOME MATHEMATICAL DERIVATIONS

Characterize U by its distribution function $F(u) = \Pr\{U \le u\}$ and its generating function $G(s)$, which is the Laplace transform of $F(u)$, namely

$G(s) = \int_{-\infty}^{\infty} \exp(-us)dF(u)$. In order for $G(s)$ to be finite, it is necessary that $dF \rightarrow 0$ very rapidly as $u \rightarrow -\infty$. Then, invariance-up-to-scale is expressed by the following conditions:

Weighted Mixture. It is necessary that stability hold for equal p_n. In particular, it is necessary that the function F satisfy the condition that

$$\frac{1}{N} \sum F\left(\frac{u}{a_n}\right) = F\left(\frac{u}{a_W}\right).$$

Maximization. Now, it is necessary that $F(u/a_M) = \Pi F(u/a_n)$; in other words,

$$\sum \log F\left(\frac{u}{a_n}\right) = \log F\left(\frac{u}{a_M}\right).$$

Aggregation. It is necessary that

$$\sum \log G(a_n s) = \log G(a_A s).$$

It turns out that the three types of invariance lead to "functional equations" of almost identical form, although they refer to different functions, respectively, F_W, $\log F_M$ and $\log G_A(s)$. Therefore, general solutions of these equations are alike. They assume the following forms

$$F_W(u) = C' - Cu^{-\alpha}, \ F_M(u) = \exp(-Cu^{-\alpha}), \quad \text{and } G_A(s) = \exp(-Cs^{-\alpha}).$$

One easily verifies that $a_M^\alpha = a_A^\alpha = \Sigma a_{n^\alpha}$ and $a_W^\alpha = (1/N)\Sigma a_n^\alpha$.

I shall now show that the above conditions are not sufficient, and that additional requirements must be imposed upon C', C and α.

Maximization. The distribution function of a random variable must be non-decreasing such that $F_M(\infty) = 1$. This requires that $C > 0$ and $\alpha > 0$, which leaves us with $F_M(u) = \exp(-Cu^{-\alpha})$.

Mixing. In order that $F_W(u)$ be non-decreasing and satisfy $F_W(\infty) = 1$, it is necessary that $C' = 1$, $\alpha > 0$ and $C > 0$.

Aggregation. In order that $G_A(s)$ be a generating function, one can show that it is necessary that $0 < \alpha < 1$ with $C < 0$ or $1 < \alpha \leq 2$ with $C > 0$.

ACKNOWLEDGEMENT

This work was partly supported by the Office of Naval Research, under contract number NONR 3775 (00), NR-047040. I am indebted to many friends – particularly Merton Miller and Lester Telser – for detailed discussions of earlier oral and written versions.

&&&&& POST-PUBLICATION APPENDIX &&&&&

THREE ASPECTS OF THE NOTION OF RENORMALIZATION

1. Footnote 4 in the original, and comment. The many footnotes in the original, except one, were easily integrated in the text. But Footnote 4 did not fit, and it cried out to be emphasized, because it was an early allusion to the theme of self-similarity that came to dominate my life and led to fractals. This footnote 4 read as follows:

"The various criteria of invariance used by physicists are somewhat different in principle from those I propose in economics. For example, the principle of relativity was not introduced to explain a complicated empirical relation, such as scaling. I am indebted to Harrison White for suggesting that I should stress the nuances between my methods and those of physics."

Harrison White is a sociologist with a background in hard science, and his comment was made after a seminar I gave in Cambridge in 1962-3, while I was visiting as professor of economics at Harvard. At that time, little did anyone expect that 1963-4 would still find me at Harvard, having moved over from Littauer Hall to teach applied physics in Pierce Hall. This was in the right place to be reminded of a topic I had studied at Caltech in 1948, namely the 1941 "Kolmogorov" theory of turbulence. The "K41" theory concluded that the spectrum of turbulent velocity should be $k^{-5/3}$. Robert W. Stewart's group at Vancouver had been the first to observe $k^{-5/3}$ in an actual experiment, and Stewart was also visiting Pierce Hall. At this point, it became clear that my version of the method of invariances has far less to do with Einstein than with Kolmogorov.

2. The physicists' concept of "renormalization" and the economists' concept of "aggregation." Section 3 will discuss the relation between the method of invariances used in this chapter, and the physicists' *renormalization*. This last term may be unfamiliar to economists, but is conceptually close to the notion of *economic aggregation*. The latter addresses the question of how, starting with the economic rules that apply to individuals, one can obtain rules relative to families and larger aggregates. It may be, but one cannot be sure, that colleagues' interests in this aggregation helped inspire me to ask how the rules relative to daily price change can be transformed into rules relative to price change over weeks and longer periods.

3. Annotation for the physicists. Only a few years after the events described in Section 1, reporting on the original Footnote 4, a current in the mainstream of "physics" turned very successfully, to the study of the "critical points" of thermodynamics. In the resulting intellectual context, the main themes of this paper are very easy to introduce.

The scaling distribution is known in physics as a "power-law" or "algebraic" distribution.

The operations with respect to which the tail of the scaling distribution is invariant are known in physics as "renormalizations." Three different renormalization are used in this chapter, one linear and two non-linear ones. Each has its own "fixed point," namely, its own "exactly renormalizable" distribution. Therefore, the key fact of this chapter may be described as reporting a property of the asymptotically scaling distribution: it is "asymptotically renormalizable" in three different ways.

Given that this paper was written during the years preceding the original publication in 1963, it could in no way be affected by the later development that introduced renormalization into physics proper.