J. M. Berger
B. Mandelbrot

A New Model for Error Clustering in Telephone Circuits

Abstract: This paper proposes a new mathematical model to describe the distribution of the occurrence of errors in data transmission on telephone lines. We suggest: a) that the statistics of telephone errors can be described in terms of an error probability depending solely on the time elapsed since the last occurrence of an error; b) that the distribution of inter-error intervals can be well approximated by a law of Pareto of exponent less than one; the relative number of errors and the equivocation tend, therefore, to zero as the length of the message is increased. The validity of those concepts is demonstrated with the aid of experimental data obtained from the German telephone network. Further consequences, refinements, and uses of the model are described in the body of the paper.

1. Introduction: The competing models

The model that we shall propose for describing the occurrence of errors is best described by contrasting it with two earlier alternatives. *The first is the binary symmetric channel without memory.* In that case, the transmission of any given bit is not influenced by the correctness of the transmission of earlier parts of the message. It follows that the distribution of the inter-error interval is such that $Pr(t) = (1 - p)p^{t-1}$, which is the geometric distribution, the variant of the exponential law relative to integer variables; the number of errors in a sequence of N bits is a binomial random variable. The many tests performed in recent years yielding error data on actual channels, however, have demonstrated the inadequacy of this model. The error data have been qualitatively described as appearing to be comprised of bursts of errors or, in fact, bursts of bursts of errors in addition to single, independent error events.

A second type of model explains this clustering of the errors by postulating that the channel has two or more states or levels of error susceptibility, with transition probabilities between these states. Within each state the occurrence of errors is described in terms similar to the binary symmetric channel, and the probability of error is dependent upon the state. For example (Ref. 1), there will be a good state with a zero probability of an error occurring, a bad state with a large probability of error, and transition probabilities corresponding to the changes from good to bad states and vice versa. The bad state gives rise to the cluster of errors, and the transition probability is chosen to account for the occurrence of these clusters. This second type of model obviously gives rise to the qualitative features ascribed to the data, but it has been found that the simple models do not give good detailed agreement and recourse has been made to more complicated models with several states.

The model we propose in the next section is similarly motivated by the qualitative appearance of the data. It had indeed been observed (see Footnote 1 and Refs. 2 to 11) that clustering of events as well as long periods without the appearance of events were characteristic of certain processes governed by distributions of the Pareto type. The simplicity of these distributions warranted their investigation for this application.

2. A new alternative model for telephonic transmission circuits

- *Statement of the fundamental postulates*

Our alternative explanation of error bunching incorporates three distinct postulates:

(A) We shall claim that, although the distance $T_{n+1} - t_n$ between successive errors is surely not ruled by a geometric distribution the random variable $T_{n+1} - t_n$ is *statistically independent* of the numbers t_i which specify the positions of earlier errors ($i < n$). The reader will observe that random variables will always be designated by capital letters, while their possible values are denoted by the corresponding lower-case letters; however, comparatively little harm will be incurred in the present case if the reader disregards this distinction.

(B) We shall assume that the distribution $F(t) = Pr(T_n - t_{n-1} < t)$ *attributes to small values of t a probability very much superior* to that corresponding to the geometric distribution.

(C) We shall also assume that $F(t)$ *attributes to large values of t a probability very much superior* to that corresponding to the geometric distribution. As a result, medium-sized t will have an unusually small probability.

A simple way of satisfying (B) and (C) is to assume that $F(t)$ is a Paretian random variable, which means that

(B′) $1 - F(t) = t^{-\alpha'}$ for small t

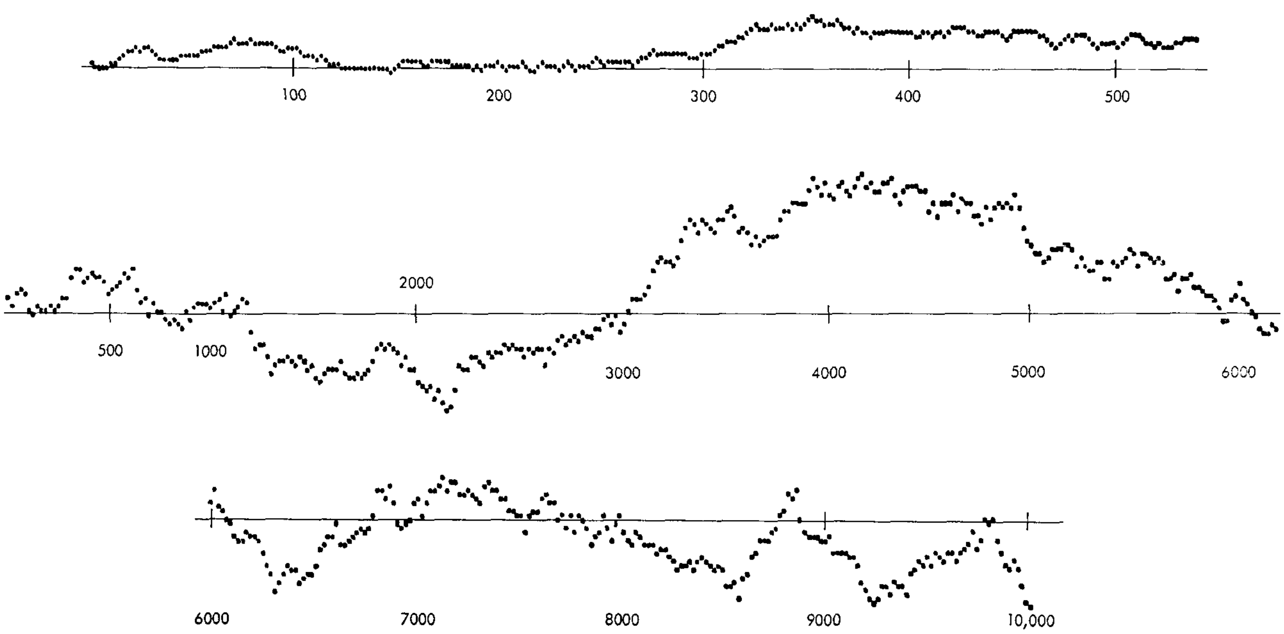
(C′) $1 - F(t) = Pt^{-\alpha''}$ for large t (Footnote 2).

In general, α' and α'' could be two different positive numbers. Things would, of course, be greatly simplified if $\alpha' = \alpha''$ (implying that $P = 1$), but the general idea of our model does not depend upon this equality. In fact, our different assumptions need not even be satisfied with equal degrees of precision.

• *Motivation of the fundamental postulates by the random walk process*

The intuitive reason that suggested the postulates (A) and (C′) is the striking qualitative resemblance which seemed to us to exist between the empirical records of inter-error intervals and the sequences of returns to the origin in the classical game of tossing a fair coin. Let us restate the rules of that game. The two celebrated old men Peter and Paul began to play (circa 1700 A.D.) with infinite fortunes; whenever their coin fell on "heads," Peter paid a cent to Paul, and whenever it fell on "tails," it was Paul who paid a cent to Peter. The behavior of $G(m)$, Peter's gain after m coin tosses, is well known to mathematicians and to some professional gamblers to be totally contrary to what is sometimes referred to as "intuition." Examine indeed our Fig. 1 (which is reproduced from Fig. III. 5 of Ref. 13). By definition, the intervals between successive roots of the equation $G(m) = 0$ are given by independent random variables; there is no question, however, that they appear to be grouped in clusters and that there are violent fluctuations in the intervals between such roots (Footnote 3). This suggests that error clustering and the violent fluctuations in the bit-error rate of telephone lines *need not* be due to dependence between the inter-error intervals; both perhaps may be described by peculiarities of the distribution of *independent* successive inter-error intervals.

Figure 1 **Record of Paul's winnings in a coin-tossing game, played with a fair coin.** *Zero-crossings appear strongly clustered, although the intervals between them are obviously statistically independent. In order to appreciate fully this Figure, one must note that the unit of time used on the second and third lines equals 20 plays. Hence, the second and third lines lack detail and each of the corresponding zero-crossings is actually a cluster or a cluster of clusters. For example, the details of the clusters around time 200 can be clearly read on line 1, which uses a unit of time equal 2. The present graph is reproduced from Fig. III.5, Ref. 13.*



Returning to the distance between successive roots of $G(m) = 0$, it is well known that its probability is nonzero only if m is even and is then equal to

$$\binom{1/2}{m/2}(-1)^{(m/2)-1} \sim \text{constant}, \; m^{-3/2},$$

which asymptotically follows Pareto's law with $\alpha = 1/2$. We took this Paretian behavior seriously, but not the precise value of $1/2$ for the exponent α, and we were thus led to our assumption (C′). As to assumption (B′), it was added in order to account for the possibility that transmission errors be even more clustered than the roots of the function $G(m)$ of coin tossing.

The novelty of our suggestions is that *they show how to generate apparent patterns of "contagion" by the choice of a process of independent events; i.e., they generate nonstationary sequences with the help of stationary processes* (for many similar examples, see Refs. 2 to 11).

Further illustration of the clustering properties of Paretian distributions

Our suggestion is further motivated by Fig. 2, fully explained in Reference 3. Let us consider three successive errors and let us suppose that the positions t_{n-1} and t_{n+1} are known, while T_n is not; the distribution of the random T_n will be studied under three basic assumptions.

When the inter-error distances are *geometrically* distributed, the distribution of T_n is *uniformly* distributed between t_{n-1} and t_{n+1}, so that the actual instants t_i are neither uniform nor bunched.

If the inter-error intervals followed a *binomial* law (a situation not encountered in practice), they could be approximated by a Gaussian, and T_n would also be a *Gaussian* variable, having $(1/2)(t_{n-1} + t_{n+1})$ as its mean value. Errors would tend to be almost uniformly distributed.

But let $T_{n+1} - T_n$ be a Paretian variable. In that case, the probability of a value t' of $T' = T_n - T_{n-1}$, given the value t of $T = T' + T'' = T_{n+1} - T_{n-1}$, is closely approximated by

$$\frac{\alpha^2 t'^{-(\alpha+1)}(t - t')^{-(\alpha+1)}}{2\alpha t^{-(\alpha+1)}} \sim (\alpha/2)t^{-(\alpha+1)} \times \left[\frac{t'}{t}\left(1 - \frac{t'}{t}\right)\right]^{-(\alpha+1)}.$$

In other words, the distribution of T_n will have *two very sharp peaks,* located at the instants of time $t_{n-1} + 1$ and $t_{n+1} - 1$, and having equal amplitudes independent of t. As a result, the middle error will "huddle" with either of the end errors, thus creating a cluster of two.

Similarly, in the case of a large number N of errors, a sizeable portion of the total sample length will be found in a few of the longest error-intervals, say in L of them, thus creating a pattern in which errors are mostly grouped in L clusters.

It will naturally be desirable to "explain" the preceding phenomena by reducing them to seemingly more elementary physical facts. We shall not attempt to do so in this paper and shall only check the model and develop its consequences. Our approach is "phenomenological", in the sense generally used to describe classical thermodynamics.[6]

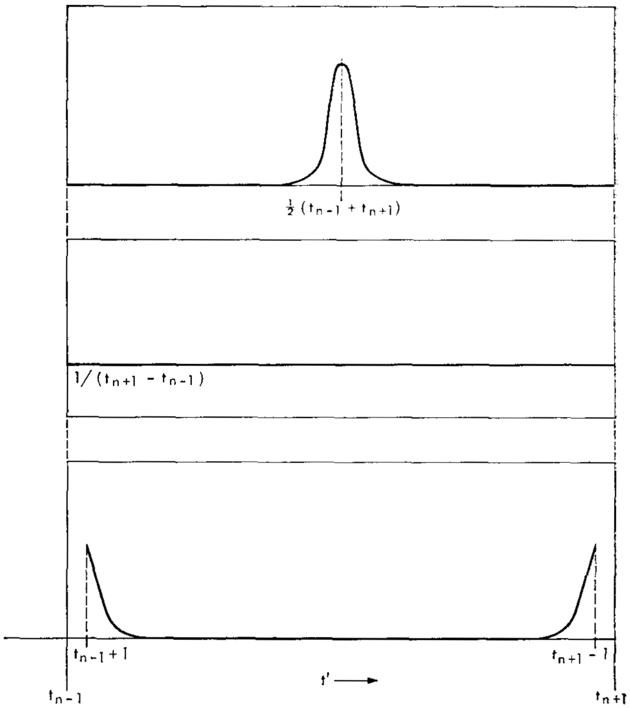


Figure 2 **Three cases of the distribution of the inter-error interval $T' = T_n - T_{n-1}$, when successive intervals $T_n - T_{n-1}$ and $T_{n+1} - T_n$ are statistically independent, and when the value t of $T = T_{n+1} - T_{n-1}$ is known.** *In each case, the horizontal scale is that of t', the vertical scale is that of the probability of t'. (For the sake of legibility, this probability is interpolated by continuous lines; the reader should recall that the geometric distribution is nothing but a discrete form of the exponential.)*

3. First experimental verification of the new model

Paretian doubly logarithmic plots

The standard method of checking the law of Pareto for $T_{n+1} - T_n$ consists in plotting $\log Pr(T_{n+1} - T_n \geq t)$ as a function of $\log t$, for the available data, in order to determine how well the resulting graph is approximated by a straight line. Unfortunately, this procedure has been neglected by statisticians, so that no small-sample tests are available and valid conclusions can be drawn only in the case of very long samples. The corresponding most natural test of the independence of $T_{n+1} - T_n$ with respect to the past will consist in plotting $\log Pr(T_{n+1} - T_n \geq t)$ *for every pattern of past errors.* Unfortunately, again, the need for very large samples has limited us to the first-order transitions and some of high order. Therefore, in varying t' from 1 to 100,000, we have performed the following tests.

We studied the function

$$\log Pr(T_{n+1} - T_n \geq t'', \quad \text{if} \quad T_n - T_{n-1} = t'),$$

which depends upon *three* successive errors.

Then we studied the function

$\log Pr(T_{n+1} - T_n \geq t''$, in the following cases:

$$\textit{either} \quad T_n - T_{n-1} = t', \qquad T_{n-1} - T_{n-2} = 1, \quad \text{and} \quad T_{n-2} - T_{n-3} = 1,$$

$$\textit{or} \quad T_n - T_{n-1} = t', \qquad T_{n-1} - T_{n-2} = 1, \quad \text{and} \quad T_{n-2} - T_{n-3} = 2,$$

$$\textit{or} \quad T_n - T_{n-1} = t', \qquad T_{n-1} - T_{n-2} = 2, \quad \text{and} \quad T_{n-2} - T_{n-3} = 1,$$

which depends upon *five* successive errors.

Finally, we studied the distribution of the sum of *three* successive inter-error intervals, $T_{n+1} - T_{n-2}$, in order to compare it with the distribution of the sum of three independent random variables having the same distribution as $T_{n+1} - T_n$: It is known that the Pareto graph of the sum of three independent variables is deducible from the graph of one by a vertical translation of log 3 of the portion of the curves corresponding to high values of the variable (see for example Ref. 3).

The results of various plots of these three functions are given in Figs. 3 through 6. While a more refined study of the model and the supporting empirical evidence will be made in Section 5, it is appropriate to examine here the more general conclusions that may be drawn from these Figures. The data used for these presentations are the part obtained from the tests described in Appendix I that corresponded to the transmit level of -22 dbm. The approximate over-all sample size is indicated by reference to Table I of that Appendix which describes similar tests. Thus, approximately 8×10^7 bits were transmitted at this level and our sample of error-bit events is of the order of 10^4. Therefore, the sample size for the marginal distributions shown in Figs. 3 and 4 is of the order of 10^4.

These plots of the marginal distribution $Pr(T_{n+1} - T_n \geq t'')$ constitute the first test of the model. The fact that a straight line with slope of about 0.4 fits this curve very well over approximately four decades of inter-error intervals is considered to represent a definite first-order verification of that portion of the model. The behavior in the last decade is not considered too meaningful for reasons that will be discussed in Section 4.

The plots of the first of the functions listed, $Pr(T_{n+1} - T_n \geq t''$, if $T_n - T_{n-1} = t')$ are given in Figs. 3 through 6 for various values of t'. The sample size is a decreasing function of t' and is of the order of 5000 events for $t' = 1$, 2000 for $t' = 2$, 1000 for $t' = 3$, 500 for $t' = 4$, 500 for t' between 10 and 20, 100 for t' between 50 and 60, and 300 for t' between 100 and 200. If the inter-error intervals were independent, we would expect to find the same distribution for all values t'. The shapes of the curves are indeed strikingly similar and, perhaps apart from the case of $t' = 1$ which is discussed further in Section 5, indicate at least a first-order verification of the hypothesis of independence. The ordering of the curves for $t' = 2, 3, 4$ does not appear to us to be significant at this time.

The plots of the second of the functions are similar to the first except that they single out some specific previous patterns which correspond to typical "bursts" of errors as discussed in the literature. These plots given in Figs. 5 and 6 are necessarily based on small sample sizes of the order of a hundred events. Since they exhibit the same characteristics as the other curves, they further establish the hypothesis of independence and our thesis that clusters *per se* or "bursts" do not have a separate intrinsic meaning and identity.

The third function, $Pr(T_{n+3} - T_n \geq t'')$, plotted in Fig. 3, is based on a sample of several thousand events. It is seen that, consistent with the assumption of independence, the vertical displacement by log 3 yields good agreement.

These results, *in toto*, appear to establish quite definitely the main features of the proposed model. The geometric distribution, which is plotted on Fig. 7 for several values of its parameter, is certainly inapplicable.

- *Joint distribution of successive inter-error intervals*

Pareto's graphical method has been criticized on various grounds. Actually, if the alternative hypothesis is the geometric distribution, and if alpha does not exceed (say) the value 3, this procedure is much better than is sometimes believed (see Fig. 7 and Ref. 6). It seems useful, however, to present another form of the evidence, originating in Refs. 5 and 10.

Let us begin by considering continuous variables, T' and T'', such that $Pr(T' > t') = t'^{-\alpha}$ for $t' > 1$ and $Pr(T'' > t'') = t''^{-\alpha}$ for $t'' > 1$. If they are independent, their joint probability density will be given by the product of the marginal densities:

$$\alpha t'^{-(\alpha+1)} \cdot \alpha t''^{-(\alpha+1)} = \alpha^2 (t't'')^{-(\alpha+1)}.$$

Hence, the lines of equal density will be given by $t't'' =$ constant: they are hyperbolas truncated to the region where both t' and $t'' > 1$.

This is to be contrasted with the two usual cases. When T' and T'' are both exponential, the lines of equal probability are straight and parallel to the line $t' + t'' = 0$. When T' and T'' are Gaussian, the lines of equal probability are circles. The contrast between these three types of curvature is so sharp that no elaborate "goodness of fit" procedure is necessary to decide which of them gives a better representation of the data relative to a very large sample.

The cases of integer-valued variables T' and T'' are very similar but more complicated to write down analytically.

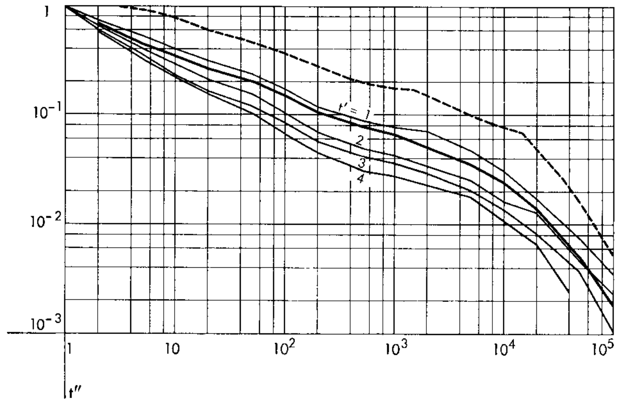


Figure 3 **Cumulated doubly logarithmic plots of empirical inter-error distributions at the transmit level of −22 dbm.** *Bold line: marginal (unconditioned) frequencies* $Pr(T_{n+1} - T_n \geq t'')$. *Dashed line: marginal frequencies* $Pr(T_{n+3} - T_n \geq t'')$. *Thin lines (looking from the top down): conditioned frequencies* $Pr(T_{n+1} - T_n \geq t''$, *when* $T_n - T_{n-1} = t')$, *for the following values:* $t' = 1$, $t' = 2$, $t' = 3$, $t' = 4$. *The peculiar behavior for* $t' = 1$ *is discussed in the body of the paper.*

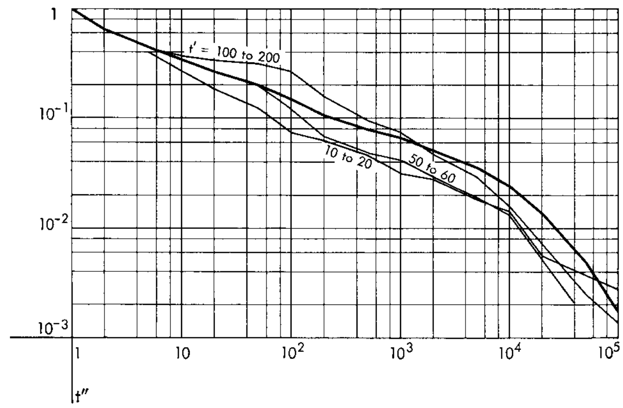


Figure 4 **Cumulated doubly logarithmic plots of empirical inter-error distribution at the transmit level of −22 dbm.** *Bold line: same as in Fig. 3. Thin lines (looking from the top down, in the region of* $t'' = 100$*): conditioned frequencies corresponding to the following ranges of values of* t' : t' *between 100 and 200,* t' *between 50 and 60,* t' *between 10 and 20.*

Figure 5 **Cumulated doubly logarithmic graphs of the following empirical inter-error distributions:** *thin line:* $Pr(T_{n+1} - T_n \geq t''$, *when* $T_n - T_{n-1} = 3$, *irrespectively of the positions of still earlier errors): dashed line:* $Pr(T_{n+1} - T_n \geq t''$, *when* $T_n - T_{n-1} = 3$ *and either of the following is true:* $T_{n-1} - T_{n-2} = 1$ *and* $T_{n-2} - T_{n-3} = 1$, *or* $T_{n-1} - T_{n-2} = 1$ *and* $T_{n-2} - T_{n-3} = 2$, *or* $T_{n-1} - T_{n-2} = 2$ *and* $T_{n-2} - T_{n-3} = 1)$. *Granted that the second curve is based upon a much smaller sample, the difference is negligible. Extremely similar results hold for* $T_n - T_{n-1} = 2$ *or 4.*

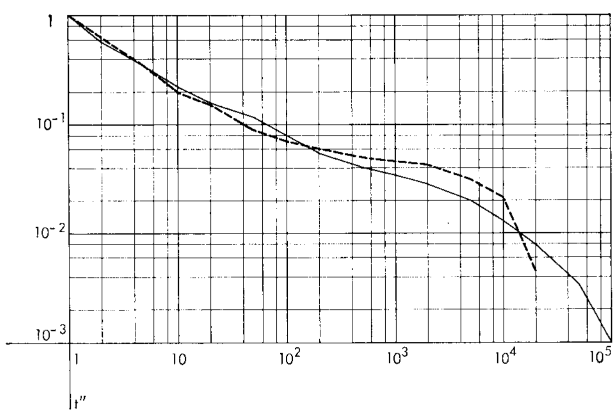
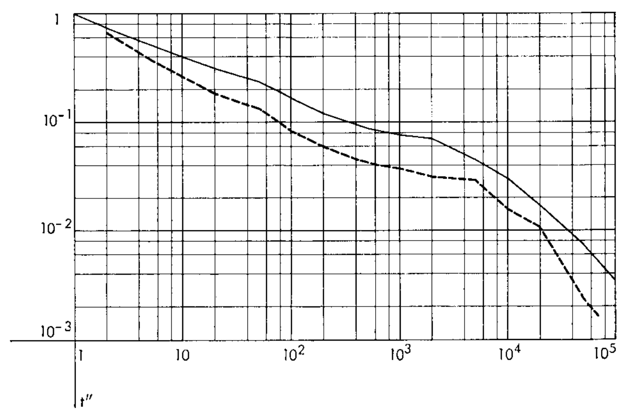


Figure 6 **Data similar to those of Fig. 5, except that** $T_n - T_{n-1} = 1$ **throughout, instead of** *3. The lower curve is very close to the three curves of Fig. 3 that correspond to* $t' = 2$, *3 or 4, and therefore to both curves of Fig. 5. This seems to mean that there is no need here for the second-order correction which is suggested by the special behavior of the line* $t' = 1$ *of Fig. 3.*



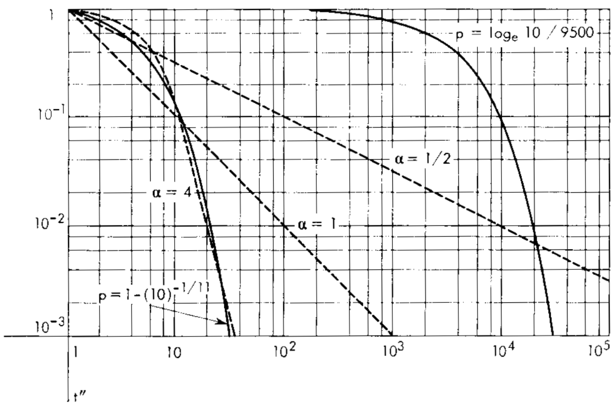


Figure 7 **Cumulated doubly logarithmic plots of two geometric inter-error distributions, corresponding respectively to a high and low value for the probability p of the mutually independent errors.** *The dashed lines give two Paretian laws of exponent, ½ and 1, as well as a law that becomes Paretian with exponent 4 when t'' exceeds 10.*

Figure 8 reproduces a three-dimensional model of the joint probability distribution. Although the details are difficult to perceive in the photograph, the display clearly favors the Paretian hypothesis, as against the Gaussian and exponential alternatives.

4. Some consequences of the new model of transmission errors

In Section 5, we shall see that higher-order complications must be introduced in order to account fully for the data. This will not change the general implications of the model, but will make them somewhat less clear-cut. It is therefore proper to interrupt the examination of the empirical verification at this stage, in order to draw some implications in the clearest case.

The number of finite moments of the inter-error interval

The usual methods of statistics have been designed for the study of random variables possessing finite moments of all orders, or at least up to order 2. However, Paretian variables with $0 < \alpha < 1$ (such as those encountered in coin tossing) have no finite *population moment* of *any* order, since the k^{th} moment is given by a divergent sum of the form

$$\alpha \sum_{t=1}^{\infty} t^{k-1-\alpha}.$$

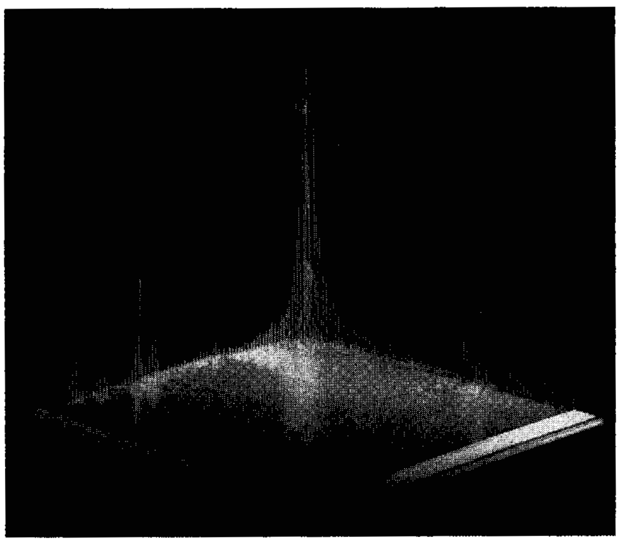
If $1 < \alpha < 2$, the first moment is the only finite one. At the same time, when the sample length is fixed, the *sample moments* of all orders are finite. The useful consequences of infinite population moments are therefore those relative to the *variation* of sample moments with sample size (Appendix II). For example, the theory predicts that, in the case of Paretian random variables with $0 < \alpha < 1$, the first sample moment will vary quite erratically with sample size and will, on the whole, tend to increase like the power (sample size)$^{-1+1/\alpha}$.

Unfortunately, this phenomenon cannot be illustrated here with the help of our error data, which—as we shall see—have been unintentionally censored of the stretches most important in our context: those containing very *few* errors.

Despite this circumstance, we see that it is very important to determine the value of α. But we originally

Figure 8 **Photographic reproduction of a three-dimensional model of the joint probability distribution of $T'' = T_{n+1} - T_n$ and** $T' = T_n - T_{n-1}$. *Actually this model was based upon incomplete data, in which several transmit levels had been mixed together. However, given the lack of precision of this photograph, there was no need to make a new model for the level of —22 dbm.*

Note concerning the coordinate scales: we wanted to fit the entire display within reasonable limits, while preserving the legibility of the most important part, that close to the origin. For that, we have used variable coordinate scales, in which successive units correspond to the following intervals of t' and t'': units from 1 to 49, tens from 50 to 99, hundreds from 100 to 999, etc. . . . To each of the changes of scale corresponds a kind of "spike," which has no intrinsic meaning.



found that the data were disappointing from the viewpoint of the sign of $\alpha - 1$. Indeed, judging from error-intervals of up to 10,000, α was markedly small (1/3 to 1/2), and even tended to decrease with $1/t$. But the local slope of the Pareto graph suddenly increases when t passes 10,000, and it may even exceed 1 when t equals 100,000. This fact did not, however, lead us to abandon outright the conjecture that alpha is the same for all t and of the order of 0.4. We argued that it is very likely indeed that, for some connections at least, our new mechanism is combined with independent errors having a very small probability of occurrence. If so, the observed curve should be expected to be some average of a straight line and of the curve corresponding to the geometric distribution as plotted on a doubly logarithmic graph; this is indeed how the empirical curves look.

Recently, however, our lack of concern with the final sections of our curves has been sanctioned in a better way. We found indeed that all runs including fewer than 5 errors were not represented on our tape, because they had been felt to be statistically irrelevant. This means that the actual data have more "tails" than the curves which we have plotted. We wish even to conjecture that the correct value of alpha corresponds to the linear extrapolation of the flattest part of our curves, i.e., that it is closer to 1/4. A larger alpha would then be required to represent short inter-error intervals.

In any event, we feel confident that there is sense in drawing the consequences of the assumption that $\alpha < 1$.

- *The number of errors per m symbols,* N_m

It has long been known in coin-tossing (see Ref. 13, p. 83) that if a play is made of m tosses, the number of roots of $G(m) = 0$ is given by one-half of the Gaussian distribution:

$$Pr(N_n < xm^{1/2}) = (2/\pi)^{1/2} \int_0^x \exp\left(-\tfrac{1}{2}s^2\right) ds.$$

Roughly speaking, this N_m will therefore increase as $m^{1/2}$, even though the most probable single value for N_m always remains equal to 1. More generally, the following was proved in Ref. 14: Let the known quantity m be the sum of an unknown number N_m of identically distributed random variables all following a law of Pareto with $0 < \alpha < 1$. Then, for large m, $N_m m^{-\alpha}$ will tend towards a nondegenerate limit: That is, the mean of N_m is $[\sin(\alpha\pi)] \cdot (\alpha\pi)^{-1} m^\alpha$, the standard deviation of N_m also increases as m^α, and the relative number of errors N_m/m will almost surely tend to zero with $1/m$.

Applying this result to the present subject-matter, we see that *the average number of digits in error should be expected to tend to zero as the length m of the message increases to infinity.*

It follows that the equivocation of our channel is zero. Indeed, U_m and V_m being an emitted and received sequence of m symbols, Shannon has defined the equivocation per sequence as being equal to $-\log Pr(U_m / V_m)$. In general, the precise definition of this conditional probability raises difficulties in the case of channels with infinite memory. In the present case, however, equivocation is simply the nonaveraged information required to specify the positions of the digits in error, i.e., the distances between the beginning of the sequence and the first error and the distances between succeeding errors.

Writing $p(t) = Pr$ (inter-error interval $= t$), we see that the specification of an inter-error interval requires on the average the information

$$\sum_{t=1}^{\infty} p(t) \log_2 p(t) \sim \log_2 \alpha - (\alpha + 1) \sum_{t=1}^{\infty} t^{-(\alpha+1)} \log_2 t,$$

which is a finite quantity Q. Moreover, the information required to specify the position of the first error is surely contained between 1 and $\log_2 m$. Hence, averaging the equivocation with respect to all possible positions of the errors, we find it to be in the interval

$$(N_m - 1)Q \quad \text{to} \quad (N_m - 1)Q + \log_2 m.$$

Finally, averaging with respect to values of N_m, we find for large m that the averaged equivocation is of the form "constant$\cdot m^\alpha \cdot Q$." The constant alone will be modified by the residual dependence to be examined in Section 5. Hence, the equivocation per symbol tends to zero with $1/m$, and, despite the presence of an unbounded number of errors, *the channel which we have defined has a capacity equal to one.*

The limit theorems of the theory of information are therefore inapplicable to binary transmission over actual telephone channels.

Generally speaking, the mathematical theory of coding, that is, the theory of information as understood in the strictest sense, consists in evaluating the various "precorrecting codes" suggested by inventors, and in comparing them with an ideal of performance associated with certain probability limit theorems. The theory of information was divided by Shannon into two parts, according to the absence or the presence of noise.

Actually (Footnote 4), this division is somewhat of an oversimplification, because the theory of noiseless transmission is *not* the limit of the theory of transmission in the presence of vanishingly small noise. For example, since the capacity of the circuit of Section 3 is equal to one, there seems to be no need for error correction of any kind. But, actually, as the word length increases, the limit of the best error-precorrecting code is not identical to the best code corresponding to the limit capacity of one. This type of limit behavior is frequently observed in engineering, and is referred to as being a "singular perturbation"; its classical prototype arises in the comparison of nonviscous fluids with those of very small viscosity. Note that the inefficiency of error correction has already been pointed out in the literature (see Refs. 18 and 19).

- *On the practical utility of the new model of errors*

It is a simple matter to program a computer to generate sequences of errors according to our first-approximation model or to our second approximation; their immediate practical consequence is therefore that it becomes a simple matter to Monte Carlo the comparison between codes. Even this semiempirical method surely compares favorably with the completely empirical approach based upon the observation of actual performance, with the help of a single long taped sequence of errors, or with the help of a new sequence of errors for each coding scheme. Our error model may also be used in analyzing queues in a communication link, in studying networks of such data circuits, and in finding optimum block-length in error-detection and decision-feedback systems.

One should not disregard, however, the difficulty of estimating the parameters of our model, at least until statisticians develop adequate small-sample techniques.

In addition, even though the present model provides no standards of optimal error precorrection, there is good hope of solving the classical probability problems required to compare the performance of various precorrection schemes when the noise follows our model.

5. Further experimental and theoretical considerations

• *Other experiments and some amendments to our model*

The initial model, involving a single parameter alpha, does not account for the detailed structure of our various Figures. Of course, we think that the model of Section 2 is a workable first approximation. For example, the conditioned distributions of Figs. 4 and 5 differ from the marginal distribution by a factor of two at most—while the geometric distribution is off by a factor of over 1000. However, it would be good to account for those variations and also to include in the model the fact that the density of errors seems to increase as the level of the signal decreases. We suggest, therefore, that the data can be better represented by either of the following two-parameter laws.

The first law supposes that the inter-error distances is such that

$$Pr(T_{n+1} - T_n = 1) = 1 - p$$

$$Pr(T_{n+1} - T_n \geq t, \quad \text{knowing that} \quad t \geq 2) = p(\tfrac{1}{2}t)^{-\alpha}.$$

Here the additional parameter is the probability that the inter-error interval be equal to 1, and it may be made dependent upon past errors.

The second possibility is to write

$$Pr(T_{n+1} - T_n \geq t) = \left(\frac{t - V}{1 - V}\right)^{-\alpha}.$$

This V was extensively used in the theory of word frequencies (Ref. 20); it could also be made dependent upon the positions of past errors.

Indeed, tests similar to those of Section 3 have been performed in the case of transmit levels of -10 dbm, -16 dbm and -28 dbm (see Appendix I). Some of the results are plotted on Figs. 9 through 12. The general appearance of these graphs is the same as for the transmit level of -22 dbm, but Figs. 11 and 12 suggest that the parameters alpha and p or V must be made to depend upon the outside conditions. We shall not study the dependence of $T_{n+1} - T_n$ on transmit level any further.

We shall, however, examine in greater detail the assertions concerning $T_{n+1} - T_n$, made in Section 3. For that, we shall begin by plotting the variation of the function

$$Pr(T_{n+1} - T_n = 1 \quad \text{when} \quad T_n - T_{n-1} = t'),$$

when its argument t' varies from 1 to 10,000. It is clear in

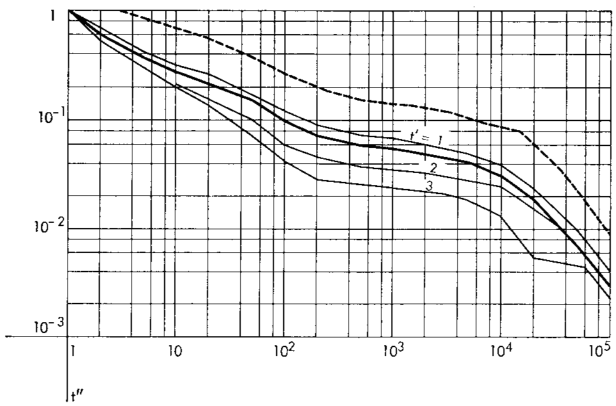
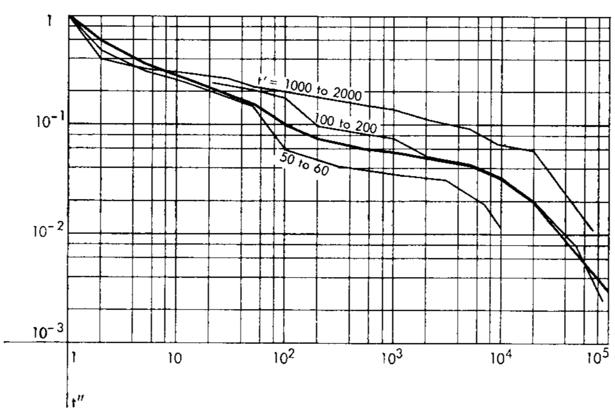


Figure 9 **Data similar to those of Fig. 3, but the transmit level was −28 dbm, and only three thin lines are given:** *$t' = 1$, $t' = 2$, and $t' = 3$.*

Figure 10 **Same explanation as for Fig. 4, but the transmit level was −28 dbm.** *Looking from the top down (in the region of $t'' = 100$), the three thin lines correspond to the following ranges: t' between 1000 and 2000, t' in the region between 100 and 200, t' in the rigion between 50 and 60.*



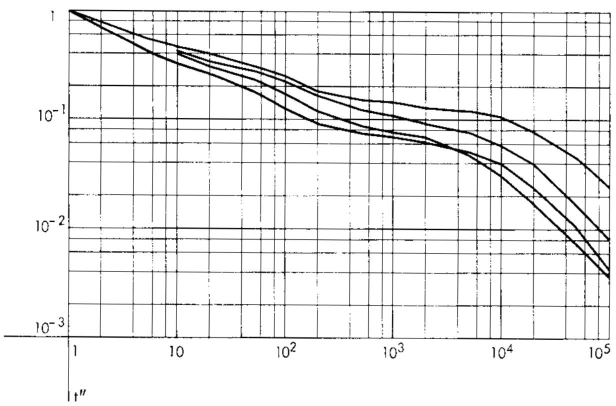


Figure 11 **Cumulated doubly logarithmic plots of the distribution** $Pr(T_{n+1} - T_n \geq t''$, **given that** $T_n - T_{n-1} = 1)$, **for four transmit levels.** *Looking from the top down (in the region around* $t'' = 100$*), the four lines correspond to the transmit levels of* -28 *dbm,* -22 *dbm,* -16 *dbm, and* -10 *dbm.*

Figure 12 **Data similar to those of Fig. 11, except that the four curves correspond to marginal (unconditioned) distributions of** $Pr(T_{n+1} - T_n \geq t'')$.

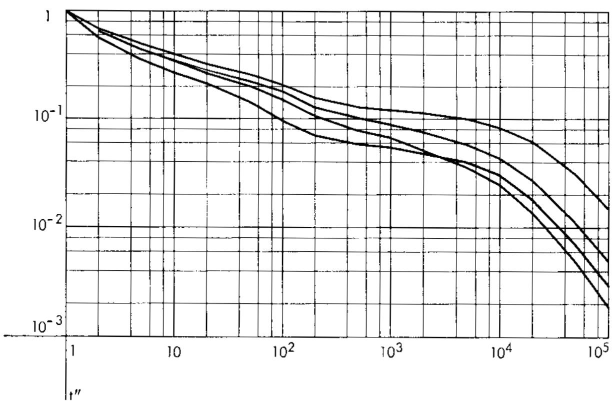


Fig. 13, based upon a sample of several thousand events, that this function is practically independent of t', except for perhaps for $t' = 2$ but mostly except for $t' = 1$, where its value is much smaller than elsewhere. This means that *sequences of three successive errors are actually markedly less frequent than predicted by our first-approximation model.* In other words, *our model is too successful in predicting the extent of clustering of errors even when the inter-error intervals are statistically independent.* The number of "clusters" of the form error-correct-error-error is also somewhat overestimated.

There is one disturbing implication in the test conditions. Depending on the length of the dialed-up connections and the character of the transmission facilities (microwave or loaded cable) and because the test was a loop test, there is a high probability of clustered-error correlation time equal to the total transit time around the loop. Thus a disturbance (e.g., dropout) at the transmitting end is likely also to disturb the currently received baud. This loop test condition is atypical of normal transmissions. It is the reverse of the error trend underlined at the preceding paragraph.

In contrast, the curve for $T_{n+1} - T_n = 2$, shown also in Fig. 13, and based on a smaller sample, indicates no systematic deviation from independence of t'.

In the same vein, it is illuminating to consider the conditional distributions of $T_{n+1} - T_n$, given the value of $t' = T_n - T_{n-1}$, and given that $T_{n+1} - T_n$ is *not* equal to 1. For that, it is sufficient to move all the conditional distributions up, until they go through the point of abscissa 2 and ordinate 1. All the various curves become practically superposable, winding around each other without clear pattern.

In other words, our earlier model may be amended, preserving an alpha independent of $T_n - T_{n-1} = t'$, while the quantity p or V introduced at the beginning of this section takes different values for $t' = 1$ and for $t' > 1$.

Our first-order model may alternatively be improved in another way, if one considers "loose clusters" for which $T_n - T_{n-1}$ is between 3 and 10. We see in the various Figures that, after such a cluster, and for t'' ranging from 50 to 10,000, $Pr(T_{n+1} - T_n \geq t'')$ is *smaller* than if $T_n - T_{n-1}$ equals 1 or 100. Hence, the observation of a "loose cluster" seems to decrease the probability that the ensuing error-interval be *very* long, even though it has very little effect for $T_{n+1} - T_n$ around 10.

- *Generation of pathological distribution by mixture of nonpathological random variables*

The pathological behaviors sketched in this paper also occur in various forms and with various intensities in the other subject matters treated in our References 2 to 11. They have repeatedly offered the suggestion that all difficulties could perhaps be attributed to the mixing of data of various origins or characters. The argument proceeds by suggesting that nonmixed data may be nonpathological. Actually, even if theoretically correct, this suggestion would not have been very useful, because it would introduce so many parameters, that reliable estimation and prediction would be impossible anyway. Moreover, the law of Pareto has some very special properties relative to mixture, as seen in Reference 6. Finally, by looking at small subsamples of our data, we have found (see Fig. 14), that differences between them are slight if the subsamples are already long, and surely nonsignificant for the small subsamples.

- *On a possible role of the model of "quality states" on a more macroscopic level*

By comparing successive very long stretches of data concerning errors (that is, stretches of several hours' duration), we found further effects that our model seems unable to explain. Therefore, even if our model represents fully the data on shorter sequences of errors, it may turn out to be necessary after all to assume that the fundamental parameters vary in time.

Acknowledgment

The authors wish to express their great debt to their colleague J. Christian of the Advanced Systems Development Division, who wrote the programs used to test our theory and generally was in charge of experimental verification.

The theoretical work reported here was supported in part by the Office of Naval Research under Contract Nonr 3775(00), NR 047040. The authors are grateful to the German Postal Administration for permission to publish the test data, which were supplied by W. Hoffman of IBM Germany.

Appendix 1 Error rates in phase modulation tests

The data that have been used in testing the proposed model were communicated to the authors by W. Hoffman of the IBM Germany Laboratories. They were obtained from tests on the German public network performed jointly by IBM Germany and the German Postal Administration. These tests were performed on both rented and dialed-up connections using both phase and frequency modulation subsets operating at 1200 bauds. Only the data obtained from the frequency modulation subset tests on the dialed-up connections have been available for our analysis. A complete description of the nature of these tests is available in Annex 12 of "Study Group Special A—Contribution No. 15" to the International Telephone and Telegraph Consultative Committee (CCITT) by the Federal Republic of Germany dated September 25, 1961. That report describes the phase modulation tests; however, the same connections and procedures were followed in the frequency modulation tests.

Briefly, four different dial-up connections were tested, each at four different transmit levels, i.e., -10 dbm, -16 dbm, -22 dbm and -28 dbm. Each level was tested continuously for 15 minutes (1.08×10^6 bits transmitted) using a periodic pseudorandomly generated 256-bit pattern. After each of the four levels had been tested, the connection was redialed so that a total of 19 hours were used in testing each of the four connections (i.e., 8.2×10^7 bits transmitted per connection). The tests were all performed on a loop basis. The returned message was compared bit-by-bit with the transmitted message and each bit error, together with its distance from the previous error, recorded on magnetic tape. This information was subsequently reduced to punched cards except for 91 of the 15-minute tests in which five or fewer bit errors occurred. The absence of these tests from our data is, if anything, a further verification of the model since their inclusion could help to decrease the curvature in the plotted distribution for very large t''.

The parameter most used for describing the data transmission performance of a communication facility is the average error rate obtained from the quotient of the total number of bit errors and the total number of bits transmitted during the test. This error rate will generally vary widely with the connection and the transmit level used. The degree of variation is indicated in the published results of the phase modulation tests which are repeated in Table 1.

The similar results for the frequency modulation data that we have used exhibit an even larger variation from level to level and between connections. It is, therefore, of considerable interest to find that the same empirical distribution as obtained from the total data fits equally well as a first approximation, independent of the transmit level and connection, and even for the limited samples of a single 15-minute test.

Figure 13 **Variation of $Pr(T_{n+1} - T_n = 1$ when $T_n - T_{n-1} = t')$, upper curve, and of $Pr(T_{n+1} - T_n = 2$ when $T_n - T_{n-1} = t')$, lower curve, for a range of values of t'.**

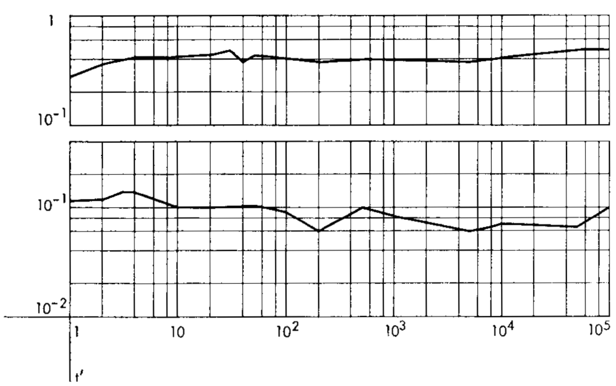


Figure 14 **Cumulated doubly logarithmic plots for several telephonic connections of short duration, considered separately.**

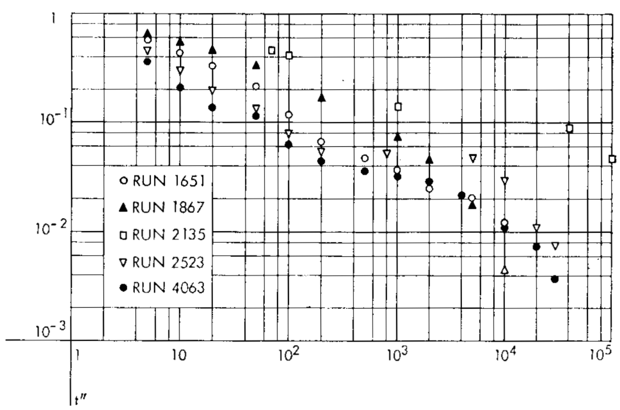


Table 1 **Total number of bits in error in phase modulation test.** *For each connection, the total number of bits transmitted per transmit level was 2.05 × 10⁷.*

Connection	*Number of bits in error transmit level* −10 *dbm*	−16 *dbm*	−22 *dbm*	−28 *dbm*
IV	162	617	2175	8430
V	513	2272	6137	15,880
VI	21	29	28	62
VII	79	417	3293	14,155

Appendix 2 Distributions with infinite population moments

We would like to reproduce from Ref. 6 some comments that elaborate upon the remarks made in the text concerning distributions with infinite population moments.

It has been said that such variables are absurd, because all empirically observed quantities being finite, their sample moments of all orders are themselves finite for finite sample sizes. This is of course true; but it does not exclude that the limits of the sample moments (corresponding to infinite samples) be themselves infinite. It has also been said that the asymptotic behavior of samples is practically irrelevant, because the sizes of all empirical samples are by nature finite. 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. All this is again perfectly true, but all that it actually implies is that the only meaningful consequences of infinite population moments are those relative to the sample moments of increasing ***subsets*** of our various bounded universes. Here, the situation is basically as follows (we shall use second moments for illustration).

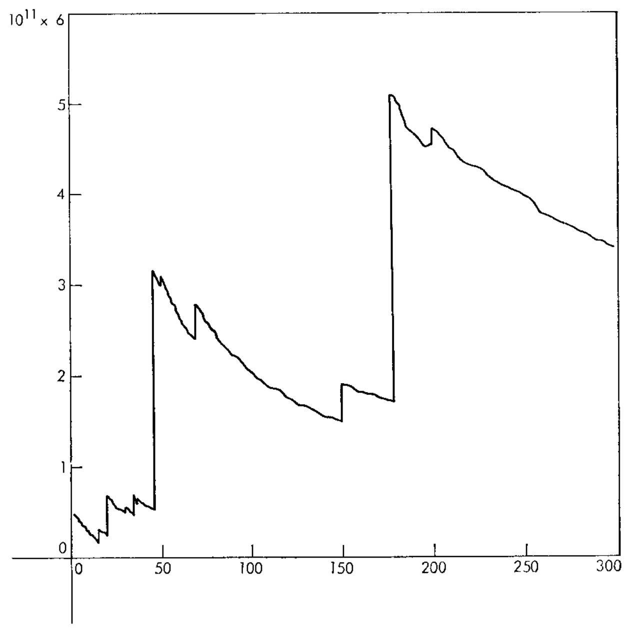
There is no question that, wherever sample second moments are empirically observed to rapidly "stabilize" around the value corresponding to the total set, it is useful to take that value as an estimate of the population moment of a conjectural infinite population, from which the sample could have been drawn. But suppose that the sample second moments corresponding to increasing subsets vary widely (Figure 15), even when the sample size approaches the maximum imposed by the subject matter. From the viewpoint of sampling, the distribution must be interpreted as such that even the largest available sample is too small for reliable estimation of the population second moment, or—in other words—that a wide range of values of the population second moment are equally compatible with the data. Insofar as the high-order moments are concerned, such an erratic behavior is almost the rule, and it is therefore prudent not to make use of them. But most statistical procedures do require first and second moments, and there it frequently turns out that the reasonable range of moment estimates happens to include the value "infinity," implying that facts can be equally well described by assuming that the "actual" moment is extremely large but finite, or by assuming that it is infinite.

In order to motivate the alternative that we prefer, let us point out that a realistic scientific model must not depend too critically upon quantities that are difficult to measure. The finite-moment model is unfortunately very sensitive to the value of the moment estimate and there are many other ways in which the first assumption, which of course is the more reasonable *a priori*, also happens to be by far the more cumbersome analytically. The second assumption, on the contrary, leads to simple analytical developments, and the rapidity of growth of the sample moment can be so adjusted that it would lead to absurd results only if one applied it to "infinite" samples, that is, if one raised problems devoid of concrete meaning. Actually, infinity is always a relative matter, entirely dependent upon 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 limit, the second moments of some variables may have to be considered as actually being finite; conversely, there are variables for which the second moment must be considered as being finite only if the useful sample size is shorter than some limit. In other words, there is no danger in assuming, as we shall do,

Figure 15 **Record of successive values of**

$$S(n) = (1/n) \sum_{m=1}^{n} U_m,$$

the sample mean of the sequence U_m, when U_m is a Paretian random variable of exponent 0.5.



that an intrinsically bounded variable is drawn at random from an infinite population of unbounded variables having infinite moments.

Actually, our use of infinity is a most common one in statistics, insofar as it concerns the function max $(u_1, u_2, \cdots u_N)$ of the observations. From this viewpoint, it would seem to be improper even to use variables with infinite spans; however, it is well known in statistics that little could be done otherwise: one even uses the Gaussian to represent the height of adult humans, which is surely positive!

From the erratic behavior of sample moments, it follows that a substantial portion of the usual methods of statistics should be expected to fail, except if extreme care is exerted. This failure has of course often been observed empirically, and has perhaps contributed to the disrepute in which many writers hold the law of Pareto; but it is clearly unfair to blame a formal expression for the complications made inevitable by the data. For Paretian laws with $2 < \alpha < 3$, second moments exist, but concepts based upon third and fourth moments, such as Pearson's measures of skewness and of kurtosis, are meaningless.

The unusual behavior of the moments of Paretian distributions can be used to introduce the least precise interpretation of the validity of the law of Pareto. For example, if the first moment is infinite, the function $1 - F(u)$ must decrease slower than $1/u$ when u tends to infinity. In this case, the behavior of $F(u)$ in the tails is very important, and, in the first approximation, it may be very useful to approximate it by the form $Cu^{-\alpha}$, with $0 < \alpha < 1$; this can never lead to harm, as long as one limits oneself to consequences that are not very sensitive to the actual value of α. If on the contrary the tail is very short (say if moments are finite up to the fourth order) the behavior of the function $F(u)$ for large u is far less important to represent than its behavior elsewhere; hence, one will risk little harm with interpolations by the Gaussian or the lognormal distribution.

Footnotes

Footnote 1. Although the validity of a specific scientific claim cannot be influenced by facts concerning other disciplines, we wish to point out that the fundamental idea of this paper was suggested by, and the techniques used are to be found in, the extensive work which one of the authors (B.M.) has devoted to "Paretian" phenomena in many different contexts. The reader who wishes to use the present work may therefore be interested in examining those references.

Footnote 2. We have discovered that our Assumption (*C'*) has already been anticipated in Ref. 12. However, Mertz appears to share the general opinion on error clustering and avoids anything resembling our Assumption (B'), by taking the law of Pareto under the truncated form $P(t) = (1 + t/h)^{-\alpha}$. Note also that many of the statistics that he uses (e.g., long-term averages) do not exist for the basic distribution and are misleading.

Footnote 3. Most books on probability, particularly the older ones, contain extensive studies of coin tossing. A useful modern reference is Feller's book.[13] The probability for the distance between roots was known at least as early as 1843. See Ref. 15. Coin tossing has been discussed in great detail by Emile Borel, and almost any of his many popular books contains many pages on this topic. These comments are addressed to the Parisian philosophers of 1900, but American engineers of 1963 should not dismiss them for that reason, since the "common sense" arguments of Borel's opponents are spontaneously re-invented by anybody who first approaches these problems. Let us quote from p. 48 of Ref. 16: "Suppose that the 2,000,000 adults living in Paris associate themselves in teams of two, and begin tomorrow morning to play at heads or tails until the winnings of both return to zero. If they work very fast, they may go through a play per second, that is, through 10,000,000 plays per year. Well, one must predict that after *10 years, 100 couples will still* be playing, and that, if the players entrust their interests to their heirs, 10 games or so will still be continuing after 1000 years."

Footnote 4. An earlier example of a singular perturbation is linked with the problem of the *limitation* of the propagation of errors that may occur on slightly noisy channels (see Ref. 17).

References

1. E. N. Gilbert, "Capacity of a Burst-Noise Channel," *Bell System Tech. J.* **39,** 1253-1265 (1961).
2. B. Mandelbrot, "Variables et processus stochastiques de Pareto-Lévy et la répartition des revenus," *Compt. rendus de l'Académie des Sciences de Paris* **249,** 613-615 and 2153-2155 (1959).
3. B. Mandelbrot, "The Pareto-Levy law and the distribution of income," *International Economic Review,* **1,** 79-106 (1960), as amended in the same *Review,* **4,** (Jan. 1963).
4. B. Mandelbrot, "Stable Paretian random processes and the multiplicative variation of income," *Econometrica,* **29,** 517-543 (1961).
5. B. Mandelbrot, "Paretian distributions and income maximization," *Quart. J. of Economics,* **76,** 57-85 (1962).
6. B. Mandelbrot, "Towards a revival of the statistical law of Pareto," to appear in *The Journal of Political Economy.*
7. B. Mandelbrot, "Very long-tailed probability distributions and the empirical distribution of city sizes," to appear in *Mathematical Approaches in the Social Sciences,* edited by F. Massarik and P. Ratoosh, Homewood (Ill.), Richard D. Irwin, 1963.
8. B. Mandelbrot, "The variation of certain speculative prices," to be published as a book (a draft, designated as IBM Research Note NC-87, was circulated in March 1962, but is now out of print).
9. B. Mandelbrot, "Geo-statistics and the law of Pareto," to appear.
10. G. L. Gerstein, and B. Mandelbrot, "Random Walk Models of the Spike Activity of a Single Neutron," Submitted to *The Biophysical Journal.*
11. B. Mandelbrot "Oligopoly, Merger, Growth and the Size Distribution of Firms," to appear.
12. P. Mertz, "Statistics of Hyperbolic Error Distributions in Data Transmission," *IRE Trans. on Communications Systems,* **CS-9,** 377-382 (1961).
13. W. Feller, *An Introduction to the Theory of Probability and its Applications,* Vol. 1, 2nd edition only, New York, J. Wiley, 1957.
14. W. Feller, "Fluctuation theory of recurrent events," *Trans. American Math. Soc.* **67,** 98-119 (1949).
15. A. Cournot, *Exposition de la Théorie des Chances et des Probabilités,* Paris, Hachette, 1843.
16. E. Borel, *Le Hasard,* Paris, Alcan, 1914.
17. B. Mandelbrot, "On Recurrent Noise-Limiting Coding," *Proceedings of the 1954 Symposium on Information*

Networks (Polytechnic Institute of Brooklyn) pp. 205-221.

18. A. B. Fontaine and R. G. Gallagher, "Error Statistics and Coding for Binary Transmission over Telephone Circuits," *Proc. IRE,* 49, 1059-1065 (1961).
19. A. Fontaine, "Applicability of Coding to Radio Teletype Channels," *MIT Lincoln Laboratories Report,* 25 G-3, October 27, 1961.
20. B. Mandelbrot, "On the Theory of Word Frequencies and on Related Markovian Models of Discourse," *Proceedings of the 1960 American Mathematical Society Symposium on the Structure of Language and its Mathematical Aspects,* edited by R. Jakobson, pp. 190-219.

Received October 11, 1962