Chapter foreword: Relations between this paper and the three main “states of randomness” described in M 1997E, Chapter E5. Most of this book deals with random functions that vary in continuous time and take continuously distributed values. In many cases, as underlined, in the book’s title, those values follow the Gaussian distribution. In contrast, this chapter deals with a point process, that is, a sequence in discrete time that equals either 0 or 1. This distribution is very far from the Gaussian.

This chapter’s original title was very different, namely “Tests of degree of word clustering in samples of written English.” This issue was of prime concern to my co-author, a linguist, and the title was geared to a journal in his field. It was a pleasure to go along because of an old interest in linguistics. M 1951 had explained satisfactorily Zipf’s law for word frequencies (see also M 1982F[FGN], starting on p. 344.) But I never ceased to wonder whether a probability can be defined for rare words. Damerau created an occasion to check on those old doubts.

A more important reason to join in this work in the 1970s – and reprint it today – lied elsewhere. I wanted to play with two statistical methods, brand new at that time and idiosyncratic, that depend heavily on graphics and the eye. To accommodate this shift in emphasis and audience, the original mathematical footnotes moved into the text and the old section on “materials and procedure” moved into an appendix.

To begin with a general issue, what is really meant by “to play with a statistical method”? I never tire of restating that every use of a mathematical method in a new context combines tests of both the context and the method. This paper tried to be open-minded about interpreting $R/S$, but
its interpretation was necessarily affected by a mental environment that preceded multifractals, therefore soon became obsolete.

To be specific, one then-new statistical method this chapter uses is $R/S$ analysis. Elsewhere in this book studies $R/S$ in the continuous context and this chapter extends it with minimal adaptation to point processes. The results yielded by this adaptation conformed to everyone's reasonable expectation. I think this suffices to make them interesting, but in a way that is far less obvious than I thought around 1970. Nice formalisms covered up a clear misunderstanding on my part, one that has no serious practical effects in this paper's context, but very serious ones in the context that will be examined in the next chapter.

In any event, using present standards, the $R/S$ tests in this paper were hasty and not detailed enough. In particular, they were limited to analysis, while synthesis had been essential to the modeling of rivers and reliefs. A reason for reprinting this text is, therefore, to encourage much-needed additional work on the use of $R/S$ for point processes.

To the best of my knowledge, the other statistical method, relying on intertoken histograms, was not developed or used anywhere else. However, it is by no means isolated in my thinking. Quite to the contrary, it fits neatly in the distinction I now make between the mild, slow and wild “states of randomness.” This line of thought is extensively explored in M 1997E (Chapter 6) and other works of mine. As a result, this work deserved being referred to in M 1997E but the thought did not occur to me in time. In a different form, intertoken intervals are also used in Hovi & al 1996.

A question of layout. To accommodate better the many illustrations in this paper, many are printed not after the first reference to them, but before.

In a natural language text of reasonable length, certain words appear to occur almost randomly while others are clustered to varying degrees. The latter category is widely believed to include content-bearing words of small overall frequency while the former category includes frequent content-bearing words and all words whose function is largely grammatical. An almost-randomly distributed word can be thought to possess a well-defined probability of occurring again in the future. As the sample of discourse increases in length, the sample frequency of a randomly distributed word is expected to converge rapidly to its probability. The probabilities of mildly clustered words are still possible to define but only as limits of more slowly converging sample fre-
quencies. As a result, longer samples are needed to estimate these probabilities with any prescribed precision. Finally, in the case of highly clustered words, the concept of word probability must be questioned and may be devoid of operational value. Extreme clustering occurs for neologisms that gain so little acceptance that the total number of occurrences is finite.

These beliefs have not been tested extensively. We performed such tests and found those beliefs to be basically sound, when appropriately tightened and hedged. Details of our conclusion are discussed later and the material, an A.P. tape and *Moby Dick*, described in the Appendix. We chose to utilize testing techniques that either are new or had not been previously used in the context of “point random processes” (of which word occurrences are an example). A desire to explore the power of such tests provided an additional motivation for this exercise.

THE STATISTICAL TESTS: PRELIMINARIES

The statistical tools utilized in this paper exemplify a difficulty statistics often encounters in the sciences. The concept of “degree of clustering” is intuitive and vague, and attempts to make it precise may lead to distinct mathematical formulations. For example, the same word type may be called clustered in one sense and not clustered in another.

To achieve perspective by an analogy, consider the history of the concept of Intelligence Quotient. Binet and the Stanford psychologists who followed had only an intuitive idea of the “intelligence” that they wanted to measure, and a great many uncertainties had to be settled more or less arbitrarily before an operational procedure implementing these ideas could be specified. Hence the claim that the Binet-Stanford I.Q. test “really measures intelligence” came to be questioned early in the development of the procedure. It appears that different I.Q.’s must be considered, each of them measuring a different “kind” of intelligence, with the original I.Q. measuring “the Binet-Stanford intelligence.”

Similarly, we will be concerned with two specific statistical techniques, both meant to assess whether a given word type is random or it exhibits “long-run clustering.” The other alternatives are “near independence of occurrences” and “short-run clustering.” Luckily, the actual classification of word types turns out to be reasonably independent of the procedure chosen. (The more usual statistical tests of independence, which we have not performed, address the extent of short-run clustering, which is an
entirely different issue.) The output of each of our tests consists of unusually many numbers printed to form a pattern. A trained statistician prefers an output consisting of very few numbers, which qualifies it as a nonredundant summary of the data. In principle, we agree with the desirability of such reduction, and we hope that ways will be found to reduce our redundant and dilute outputs. For the present purpose, however, our outputs are perfectly satisfactory because in most cases the computer output patterns that correspond to clustered and nonclustered word types differ obviously.

**RELATIVE INTERTOKEN POSITION HISTOGRAMS**

Consider the sequence of positions of the successive tokens of a word type. The position of the first token is selected as the origin of time and the subsequent positions are designated by $T_1, T_2, \ldots, T_k$, and so forth. The most important models for the distribution of the word token locations $T_k$ are the following: Figure 1.

(A) **First model.** This model is not realistic and is only useful for the sake of contrast. It assumes that the tokens are uniformly spaced and that the intertoken intervals, $T_{1'}, T_{2'} - T_{1'}, T_{3'} - T_{2'}$, and so forth, are identical. The concept of type probability is not only well-defined in this model but, in fact, degenerate since in all samples of discourse of the same large duration each token occurs precisely the same number of times. Figure 2.

(B) **Second model.** This model is not realistic either and is only used for the sake of contrast. It assumes that the tokens are statistically independent so that token positions essentially form a Poisson process. Then the intertoken intervals fluctuate. Among samples of discourse of the same duration, the relative frequency of each token fluctuates for the length of this duration. But this frequency tends eventually to a limit, making the concept of type probability, again, well-defined. The implications of the Poisson model are apparent on the distribution of the relative position of the token $T_k$ among its neighbors. First, examine the immediate neighbors $T_{k-1}$ and $T_{k+1}$, and form the ratio $(T_k - T_{k-1})/(T_{k+1} - T_{k-1})$, which can be called a “relative position of order 1.” In the Poisson case, this ratio is known to be distributed between 0 and 1 uniformly. Next, examine $h$-removed neighbors $T_{k-h}$ and $T_{k+h}$, and consider the “relative position of order $h$,” defined as the ratio $(T_k - T_{k-h})/(T_{k+h} - T_{k-h})$. In the Poisson case, the expectation of this ratio is 0.5, and its distribution is bell-shaped, clustered near the expectation. As $h$ increases, the tightness of the bell’s
**FIGURE C29-1.** Eight selected relative intertoken position histograms for the word type “several” in the AP file.

*Discussion of results.* The word type “several” is very characteristic of near independent tokens. In the case of independent tokens, the first histogram would have been completely flat. Here, it is *near* flat, except for end peaks that express the fact that the ratios starting with 0 or 9 are too numerous be due to chance. When those peaks turn out to be statistically significant, they are evidence of clustering. The question is, how strong is this clustering? If it were strong and/or global, the second histogram would also have end peaks, but in the present instance it does not. This suffices to show that the clustering was both local and slight. Its statistical significance has not been investigated further. In the case of local clustering, which includes the case of independent tokens, the last histogram is expected to be shaped like a Gaussian distribution, the “Galton ogive,” which is indeed the case.

*Conclusion.* The word type “several” has a well-defined probability. [P.S. 2000: its sample frequency exemplifies “mild” variability.]
FIGURE C29-2. Eight relative intertoken position histograms for the word type “Africa” in the AP file.

Discussion of results: These histograms are very characteristic of the case when local clustering is strong and the presence and/or intensity of long-range clustering is not clear. The presence of end peaks in several of the histograms indicate clustering. The intensity of local clustering is illustrated by the fact that the peaks in the first histogram are very high. However, as one proceeds to the right, the end peaks eventually disappear, and the last histogram has become bell-shaped. It looks different from the bell of the word type “several,” but the present technique alone is not sufficient to determine whether the difference is real. In such instances, a detailed study of long run clustering requires a different technique; for example, see Figure 9.

Conclusion. One can speak of the probability of the word type “Africa,” but the convergence of empirical frequency to this probability may be slow. [P.S. 2000: its sample frequency exemplifies “slow” variability.]
clustering near 0.5 increases; the ratio has an increasing probability of lying near 0.5. Figure 3.

The technical reason for this increasing clustering is that the intervals between independent tokens are geometric random variables. \( T_k - T_{k-h} \) is the sum of \( h \) such variables, which follows the law of large numbers. Therefore, it nearly equals \( h \) times its expectation, and, by the central limit theorem, its scatter follows the Gaussian distribution. Figure 4.

(C) **Third model.** This model makes the more realistic assumption that the token locations are statistically independent, except that neighboring words may interact. For example, tokens of the same type may be either prohibited or encouraged to follow each other closely; in the latter case, they show a slight tendency to cluster. Figure 5.

In both cases, “local” interactions allow word types to continue to have a well-defined probability. This concept only allows the departure

![Figure C29-3. Eight relative illtoken position histograms for the type “Sex” in the AP file.](image-url)

**Discussion of results.** These histograms are characteristic of near absent local clustering combined with strong long-run clustering. The most striking histogram is the last, which is not shaped like a bell but more like the letter U. Compared to the word type “Africa,” overall clustering (as seen on the end peaks of the first few histograms) is stronger. More importantly, the span of statistical dependence between the intertoken intervals is much longer.

**Conclusion.** For the word type “sex,” the notion of probability is highly controversial [P.S. 2000: its sample frequency exemplifies “wild” variability.]
from independence to affect the distributions of the relative positions of order $h$ when $h$ is small. For example, assume that the local interaction encourages clustering. Then, for small $h$, the tokens $T_{k-h}$ and $T_{k+h}$ are very likely to belong to two different clusters, and the token $T_k$ is likely to belong to one of the two clusters containing $T_{k-h}$ or $T_{k+h}$. It follows that, for small values of the order $h$, the probability distribution of the relative position is U-shaped, with maximal probability near 0 and near 1 and with minimal probability near 0.5. To the contrary, when $h$ is large, it seems reasonable to characterize the "local" character of the dependence by the requirement that the distributions of high-order relative positions become bell-shaped when $h$ is large, as in model (B) above. Figure 6.

For example, tokens can be considered as locally clustered when they can be divided into two classes: "leaders" and "followers." Leaders would follow a Poisson process as in model (B), and the number of followers of each leader would be random but not too variable. If the intertoken intervals are (a) not too far from being geometrically distributed – in particular their variance must be finite – and (b) not too far from being independent,
then the central limit theorem holds just like in model (B) but only for values of $h$ greater than the values assumed in the case of independent $T_k$.

(D) **Fourth model.** This model supposes that the token positions are as tightly clustered possible, in a single, close cluster. Then the concept of word probability becomes meaningless, but at the same time the U-shaped distribution above continues to hold even when $h$ is very large. Figure 7.

(E) **Fifth model.** The model of “long-run clustered” types is obtained by weakening the fourth model to allow more than one cluster while continuing to demand that the distributions of the relative position remain U-shaped, with maxima near 0 and 1, for every value of $h$. In other words, as $h$ increases, the distribution of the relative position must fail to converge to a bell clustered near 0.5. Figure 8.

In this fifth model, the concept of the probability of word type loses meaning, but probability ideas remain useful in a modified fashion. For example, consider the word type “sex” in the case of a printed discourse in which this word is rare and very clustered. The absolute probability

![Figure C29-5. Illustration of one extreme behavior of $R(t, \delta)$. In the case of uniformly spaced tokens, the value of $R(t, \delta)$ is equal to 1, a quantity that is small and independent of $\delta$ and of the density of tokens. $\delta^*(t, \delta)$ is clearly proportional to the density of the tokens but is independent of $\delta$. It follows that $R/S$ is independent of $\delta$ and the diagram of log $R/S$ as a function of log $\delta$ is a horizontal line.](image-url)
may well be meaningless, but, when it is known that this word type has appeared at least once in a discourse of $T$ word tokens, the conditional probabilities of its having occurred 2, 3 or more times all become meaningful. If they do, then word occurrence is ruled by a generalized random process that M 1967b[N10] introduced under the name of sporadic processes. The belief that the law of large numbers and the central limit theorem always prevail is so strong that the above definition of extreme clustering might seem logically contradictory. But in fact it is not. Figure 9.

Sample results of the relative interval analysis we have carried out are shown in Figures 1, 2 and 3, details being given in those figures’ captions. For an earlier use of an analogous technique, see Josselson 1953.

Method of construction of Figures 1, 2 and 3. In the first histogram, the abscissa is the first decimal of the ratio between an intertoken interval and

![Figure C29-6](image)

FIGURE C29-6. Illustration of a second extreme behavior of $R(t, \delta)$. In the case of $k$ tokens pressed into a single cluster, the value of $R(t, \delta)$ is approximately equal to $k$. In still another case, less extreme clustering, $R/S$ is proportional to $\delta$, and the diagram of log $R/S$ as a function of log $\delta$ is a straight line of slope 1. In all cases, the slope of log $R(t, \delta)$ versus log $\delta$ falls between the extremes of 0 and 1. More precisely, the slope of log $R/S$ versus log $\delta$ for independent tokens is a straight line of slope 0.5, and log $R(t, \delta)$ versus log $\delta$, for word tokens, is usually a straight line of slope between 0.5 and 1.
FIGURE C29-7. R/S diagram for the word type “as” in the Moby Dick sample.

Method of construction. The abscissas are values of the lag $\delta$ for which the function $R/S$ was computed. They have been selected to be spread regularly along a logarithmic scale, 10 values per decade. The ordinate axis is subdivided into “cells” bounded by values spread regularly along a logarithmic scale, 12 values per decade. (The constants 10 and 12 are due to the constraints of computer output.) For each lag $\delta$, approximately 70 values of the starting point $t$ are considered, uniformly spread along the sample, and the resulting values of $R/S$ are sorted in the above cells. The number in each cell is printed exactly when it is at most. It is represented by + when it lies between 10 and 24 and is represented by a small, filled-in circle when it is 25 or above. The median cell (defined so that less than half of the values of $R/S$ lie in cells above and below it) is underlined.

Discussion of results. This pattern is very characteristic of near independent tokens. In this case, the theory indicates that the underlined cells should, for large lags, lie along the line of equation $R/S = 1.25 \sqrt{\delta}$, which has been drawn as a straight line on this diagram. This theoretical prediction is indeed verified. For small lags, the theory indicates that the occupied cells should lie along the line of equation $R/S = \sqrt{\delta}$, whose plot is parallel to the line drawn, indeed they do.
the sum of this and the next interval. In the last histogram, the abscissa is the first decimal of the ratio between the sum of eight successive inter-
token intervals and the sum of these and the next eight intervals, and similarly for intermediate histograms. The ordinate is proportional to the number of ratios in question for which the first decimal has the value drawn as the abscissa. Figure 10.

**R/S ANALYSIS**

A second technique that we used is \( R/S \) analysis, a method of data analysis inspired by Hurst 1965. It has been formalized only recently and has been used mostly for random processes such as river discharges and commodity prices. We wanted to test this technique on a “point process,” namely, the random sequence of events constituted by word token locations. One way to handle such a sequence is to transform it into a

![Figure C29-8](image.png)

**FIGURE C29-8.** A form of the \( R/S \) diagram for the word type “several” in the AP file. The method of construction and comments are discussed in the caption of Figure 7.
function $X(t)$. Write $X(t) = 1$ if the word location $t$ in our sample is occupied by a token of the type under study; otherwise, $X(t) = 0$.

In addition, $X_u(t) = \sum_{u=1}^{t} X(u)$ is the cumulative number of tokens in the sample from $u = 1$ to $u = t$. The letters in "R/S" then denote the "rescaled bridge range" $Q(t, \delta) = R(t, \delta)/S(t, \delta)$, a function of $t$, called the "starting time," and of $\delta$, called the "lag."

In words, $R(t, \delta)$ is the "cumulated range" of a process between times $t + 1$ and $t + \delta$ after removal of the sample average, and $S^2(t, \delta)$ is the corresponding "sample variance" around the sample average. Mathematically, $R(t, \delta)$ is defined – as shown on Figure 4 – by

![FIGURE C29-9. A form of the R/S diagram for the word type “Africa” in the AP file. The method of construction is discussed in Figure 7.](image)

**Discussion.** Concerning local clustering, this diagram adds nothing to the information contained in Figure 2. But it does erase the doubts Figure 2 had left about global clustering; indeed, after $\delta = 10^3$, the line joining the underlined cells goes up with a slope greater than 0.5, which indicates that long-run clustering is present.
\[ R(t, \delta) = \max_{0 < u \leq \delta} \{ X(t + u) - X(t) - (u/\delta)[X(t + \delta) - X(t)] \} \]
\[ - \min_{0 < u \leq \delta} \{ X(t + u) - X(t) - (u/\delta)[X(t + \delta) - X(t)] \} \]

and \( S^2(t, \delta) \) is defined by
\[
S^2(t, \delta) = \delta^{-1} \sum_{u=1}^{\delta} \{ X(t + u) - \delta^{-1}[X(t + \delta) - X(t)] \}^2
= \delta^{-1} \sum_{u=1}^{\delta} X^2(t + u) - \left[ \delta^{-1} \sum_{u=1}^{\delta} X(t + u) \right]^2.
\]

In this chapter the function \( X(t) \) reduces to either 0 or 1, implying \( X^2(t) = X(t) \), therefore \( S^2(t, \delta) \) simplifies to

**FIGURE C29-10.** A form of the \( R/S \) diagram for the word type “sex” in the AP file. The method of construction is discussed in Figure 7.

*Discussion.* The diagram diverges sharply from the smooth pattern of slope 0.5 found for “several.” This confirms the results of Figure 3.
\[ S^2(t, \delta) = \frac{X_\delta(t + \delta) - X_\delta(t)}{\delta} - \left( \frac{X_\delta(t + \delta) - X_\delta(t)}{\delta} \right)^2. \]

In particular, in the study of discrete events of low frequency, as in word tokens belonging to a rare word type, one has very nearly

\[ S^2(t, \delta) = \frac{X_\delta(t + \delta) - X_\delta(t)}{\delta}. \]

In the case of a very rare word, the marginal probability distribution of the process \( X(t) \) (the distribution of \( X(t) \) disregarding the temporal ordering of its values) is extremely skew. The present work may, therefore, be viewed as extending the study of \( R/S \) analysis into the realm of very skew distributions.

\( R/S \) testing. The behavior of \( Q(t, \delta) \) as \( \delta \to \infty \) serves to define the concept of “\( R/S \) independence,” which is a form of “nonperiodic long-run statistical independence.” The first application of \( R/S \) analysis is a test of whether or not a record is \( R/S \) independent. The process of independent events (see model B above) is unquestionably the simplest point process. For model B, Feller 1951 showed that \( \lim_{\delta \to \infty} \delta^{-0.5} Q(t, \delta) \) is about 1.25, which is both positive and finite. More generally, M & Wallis 1969c[H25] has demonstrated and M 1975z[H26] has since proved mathematically that, for nearly every independent process and every process from which long-term dependence is unquestionably absent, the function \( R/S \) has the same asymptotic behavior for \( \delta \to \infty \): it satisfies the \( \delta^{0.5} \) law, which asserts that the expression \( \lim_{\delta \to \infty} \delta^{-0.5} [R(t, \delta)/S(t, \delta)] \) is well-defined, positive and finite. In intuitive terms, this means that the graph of the expectation of \( \log[R(t, \delta)/S(t, \delta)] \) versus \( \log \delta \) is asymptotically a straight line of slope 0.5 and that the scatter of empirical values around this “trend” is independent of \( s \).

A sharp contrast to this \( \delta^{0.5} \) law is found in the behavior of \( R/S \) shown in Figures 5 and 6. More interesting discrepancies occur when words exhibit unquestionable global statistical dependence, other than periodic behavior. This is a way of saying that the dependence between \( X(t) \) and \( X(t + T) \) decreases to zero as \( T \to \infty \), but does so extremely slowly. In these cases, the asymptote of the function \( \log[R(t, \delta)/S(t, \delta)] \) versus \( \log \delta \) is not a straight line of slope 0.5; either the graph is not straight or, if it is straight, its slope \( H \) is different from 0.5. When word tokens are long-term dependent but do possess a well-defined probability, the value of \( H \)
may be taken as a measure of the degree of interdependence. However, when word tokens form a sporadic process (see the end of the preceding section on relative interval histograms), the behavior is more complex. We cannot present the details necessary to discuss adequately these complications here. Our exhibits instead primarily attempt to either confirm or invalidate the hypothesis that there is long-run dependence.

Results of the $R/S$ analysis of several words are shown in Figures 7, 8, 9 and 10, whose captions contain important details.

CONCLUSION

Our examples are characteristic of other word types that we have examined, although there are occasional anomalies. Except for few clear-cut cases, the most complete analysis is a combination of the two techniques described in the paper. The addition of other techniques would undoubtedly further improve the analysis. However, a systematic classification of possible behaviors is beyond the scope of the present exercise.

APPENDIX: EXPERIMENTAL MATERIALS AND PROCEDURE

Ideally, we would want to work with texts that are both extremely homogeneous and very long. Since these requirements are to some extent contradictory, we resorted to two texts, one very long and fairly homogeneous and the other very homogeneous and fairly long. Because of the cost of preparing machine readable input, some characteristics of this large scale linguistic experiment were dictated by source material availability. In addition, for the sake of comparison, various random pseudo-texts were generated.

Our very long text, which runs slightly in excess of 1.6 million words, was generated from the Associated Press European wire and was supplied through the generosity of the Associated Press. Prior to the processing specific to this experiment, all non-English material was deleted, as well as most of the sports news, commodity and stock market reports, weather reports and the like. The remaining “news text” deals largely with the kind of events which might occupy the front page of a daily newspaper.

Our very homogeneous text, which runs slightly in excess of 118,000 words, is the whole of Herman Melville’s *Moby Dick*. A copy of this book
on punched cards was kindly provided by Professor J. Raben of Queens College.

Our samples of pseudo-texts were simulated in various ways. In some samples, it is assumed that the gaps between successive occurrences of each pseudo-word are statistically independent and follow one of several Poisson or hyperbolic distributions with different parameter values. In other samples, the gaps are dependent. The sample size was chosen to match the size of the AP news wire. We used the pseudo-random number generator recommended in Lewis, Goodman & Miller 1969.

From the AP news text, about 250 word types were selected for study. Half of the word types were chosen because of their putative membership in the class of structure words, and the other half were selected because of their putative failure to belong to this class. An attempt was made to have roughly the same frequencies for the word types in the second class as in the first class. In this selection, a word count of the first 700,000 words of the text was used as a guide. Similarly, from the Moby Dick text, about 250 words were selected for study, again on the basis of a word count.

An especially written PL/1 program decomposed the text into word tokens. When a word token in the text corresponded to one of the word types in the study list, the program recorded both the type and the location of the token in the text. After the entire sample text had been processed, these records were sorted by word type, producing a file of the location of each token occurrence of a particular word type in the text. From this file, it was a simple matter to determine the gaps between occurrences. Ultimately, both positions and gaps were plotted.

After the master files had been created, additional special PL/1 programs were written to compute and to print graphs of gap frequency, graphs of the relative intertoken position from 1 out of 2 up to 10 out of 20, and graphs for the values of $R/S$, as well as tables of the raw occurrence data.