An essay submitted in partial fulfillment of the requirements for the degree of M.S. in Statistics under the direction of Dr. E.J. Wegman.

SOME STATISTICAL TECHNIQUES FOR DISPUTED AUTHORSHIP: A CASE STUDY

Lynne Stokes

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 932

June 1974

ACKNOWLEDGEMENTS

I would like to express my appreciation to Dr. Edward Wegman for the original suggestion for this project and for his time in many summer-time meetings. Special thanks also go to Bob Hutten for his enthusiasm, his hours of consultation, and even more hours of key-punching. And I would like to thank Dr. N.L. Johnson for his conversations and suggestions in the very early stages of this project, and to Ms. June Maxwell for her patience in the very late stages.

TABLE OF CONTENTS

SOME STATISTICAL TECHNIQUES FOR DISPUTED AUTHORSHIP: A CASE STUDY

1. Introduction

For more than 80 years, attempts have been made to detect difference in literary style by mathematical methods and to "solve" cases of disputed authorship by these means. Most have been unconvincing. One obstacle to these efforts has been the problem of accumulating data, since hand counts are expensive, time-consuming, and inaccurate. T.C. Mendenhall [cf. 9] was the first to undertake a large-scale study of this kind in the late nineteenth century. A few others, such as Yule [11] and Williams [10], followed his pioneering efforts almost 50 years later. The introduction of the computer to the research scene in recent years has revived the interest in authorship studies among both statisticians and linguists since analyses of large masses of data are now feasible. Mosteller and Wallace [4] have produced by far the most detailed and mathematically sophisticated of these recent studies.

In 1887, Mendenhall [cf. 9] published a study of the frequency distributions of words of different lengths appearing in an author's writings, and he attempted to differentiate between authors on this basis. Being a physicist, he anxiously tried to produce a "word spectrum" or "characteristic curve" to identify a writer by graphing his data, but he actually did no statistical analyses of it. He did

try to give evidence that his spectra were characteristic of a writer, showing by means of examples that they persist through different samples and sample sizes involving one man's writings.

Later, he applied his method to the classic question of authorship of some works of Shakespeare, for which he counted nearly two million words. Bacon's spectrum differed markedly from Shakespeare's, but according to Mendenhall, "...Marlowe agrees with Shakespeare about as well as Shakespeare agrees with himself."[1] However, he was not so naive as to state his inference as irrefutable fact and he encouraged the reader to draw his own conclusions from the data.

Certainly there are other faults in Mendenhall's work besides his primitive statistical analyses. For example, his sampling technique is highly questionable; he described his method as taking 1000-word blocks "at the beginning of the volume and, after a few thousand words had been counted, the book was opened near the middle and the count continued."[2] As a contribution, however, Mendenhall recognized and commented that his work might be extended in countless ways by different methods and to different problems. He suggested the study of the evolution of a language or of the growth of vocabulary from childhood to adulthood as two areas to which this type of analysis might be applied. He saw one advantage of word length as a random variable in disputed authorship problems as being that the author cannot consciously control his word length effectively since he will most likely not

---

1 Williams [9], p. 252.
2 *Ibid.*, p. 253.

even be aware of it. In contrast, sentence length, the next variable seized on by statisticians, is less subtle.

A more recent word length study was undertaken by Claude Brinegar [1] in 1963 in an effort to ascertain the authorship of the "Quintus Curtius Snodgrass letters," allegedly, but questionably, by Mark Twain. He compared the word spectrum from Twain's early writing with his later style, and found the two to be quite consistent. He then: (1) performed a χ^2 test on the Twain control group and the QCS letters and (2) performed a 2-sample t-test on the proportions of 2,3, and 4-letter words. Both tests strongly rejected the theory supporting Twain's authorship.

G. Udny Yule [11] in 1938 was the first to use sentence length as a discriminating characteristic. It has the advantage that it has a greater range than word length. Yule's first goal was to show that sentence length is actually a characteristic of a writer's style. That effort consisted of tabulating sentence length data on two large samples, A and B, from each of four author's writings. He then calculated certain statistics (e.g., mean, first and third quartiles, ninth decile) for A and B and compared them simply by inspection.

Although the calculation of statistics from the samples was an improvement over Mendenhall's efforts, the method is still primitive. It should be noted, however, that some present tools were not available then to deal with such a problem. Yule has already found that the distribution was decidedly skewed, so normal tests were not appropriate. The standard two-sample, nonparametric tests were not in use at that time.

But there is a more subtle difficulty with Yule's inference that sentence length is characteristic of an author and is therefore an effective discriminator for authorship disputes. His two samples from each author are from either the same volume or are from two quite similar pieces. There is no reason to conclude that the similarity is any more indicative of the writer than of the type of writing. It seems important to show that an author's sentence length distribution is consistent throughout his career and over all topics if the disputed piece is not similar to the control group samples with respect to these characteristics.

Yule used his method to ascribe the authorship of *De Imitatione Christi* to Thomas à Kempis over Jean Charlier de Gerson. He became concerned by his sampling technique and tried a pseudo-random scheme. The data for the random and non-random samples are tabulated and described by Yule as "closely similar."

In 1939, Williams [10] took up the study of sentence length. He tried to force his data into some known distribution, presumably so that parametric tests might be applied. After manipulating the data, he decided that sentence-length follows a log-normal distribution. For his argument, he showed normal curves superimposed on his log-number-of-words data, but did not perform any goodness-of-fit test.

In 1944, Yule [12] pursued the *Imitatione* authorship study by developing a "richness of vocabulary" index. It was defined in the following manner: let $f_x \equiv$ number of words occurring x times in the sample, and $N \equiv$ sample size. Then $M_X = \sum_{x=0}^{N} f_x \, x/N$ and

$\sigma_X = \sum_{x=0}^{N} f_x x^2/N$ will increase with increasing sample size. But Yule found $K = 10{,}000(S_2/S_1^2 - 1/S_1)$, where $S_1 = N\bar{x}$ and $S_2 = N\sigma_X$ to be fairly constant with varying sample size. Using this "characteristic K", he again assigned *Imitatione* to Thomas à Kempis.

The most elaborate of all authorship studies was done in 1963 by Frederick Mosteller and David Wallace [4]. They tried many of the traditional techniques, but most were not good disciminators between Hamilton and Madison, the two contestants for authorship of certain disputed essays of the Federalist papers. They settled on two main studies; the first was Bayesian and the second was a more classical linear discriminant function. Only the first will be discussed here.

They based their analysis on the frequency of occurrence of thirty common, non-contextual words, such as "upon". These words were found to be fitted adequately with a negative binomial or even, in some cases, a Poisson distribution. The first appearance of their Bayesian spirit appears in the calculation of "log odds." Let p_1 and $p_2 = 1-p_1$ be the prior probabilities that Hypothesis 1 (Hamilton wrote the paper) is true or Hypothesis 2 (Madison wrote the paper) is true. For historians, p_1 and p_2 might take on different values, but for those with no opinions on the question, let $p_1 = p_2 = 1/2$. Let $f_1(x)$ be the conditional probability of observing x, given H_1 is true. Then, the conditional probability that H_i is true given that x was observed is given by

$$P(H_i|x) = \frac{p_i f_i(x)}{p_1 f_1(x) + p_2 f_2(x)} \quad \text{for } i=1,2 .$$

Then Mosteller and Wallace defined the "odds" of H_1 relative to H_2 by

$$\frac{P[H_1|x]}{P[H_2|x]} = \frac{p_1 f_1(x)}{p_2 f_2(x)} = \left(\frac{p_1}{p_2}\right)\left(\frac{f_1(x)}{f_2(x)}\right)$$

$$= \text{initial odds} \quad X \quad \text{likelihood ratio.}$$

They worked with log odds since then the "evidence" from independent measurements is additive. The log likelihood ratio for the Poisson distribution with parameters $\omega\mu_H$ and $\omega\mu_M$ (with $\omega \equiv$ number of words in thousands and $\mu \equiv$ word rate per thousand) is $\lambda(x) = x \log(\mu_H/\mu_M) - \omega(\mu_H-\mu_M)$.

The log odds can therefore be increased by adding the effects of many independent discriminating words. However, the decision on whether or not to include a word should consider if its contribution is worth the cost of including it. Of course, the more extreme is the value μ_H/μ_M or $\mu_H-\mu_M$, the more extreme the log likelihood ratio and thus the log odds. So words which appear often with different rates or even less often with enough discrepancy in their rates can be helpful. With the use of many and well-chosen words, the effect of the initial odds, p_1/p_2, will essentially be overwhelmed, unless, of course, $p_1 = 0$ or $p_1 = 1$.

The next question to arise is how the rate estimates are to be calculated. This is the most important use of Bayesian ideas. The complicated estimation procedure derived from their ideas is adopted to combat the "regression effects" which might arise by picking only

the words which had the most discrepant rates in the samples and using the sample rates as true rates in calculating the log odds. The data from the Federalist essays of known authorship is applied to a prior distribution, which will be described below, to obtain a posterior distribution for the word rates. For simplicity, the procedure for a Poisson rather than a negative binomial random variable will be considered.

The procedure for establishing a prior distribution was begun by introducing a pair of parameters for each word:

$$\sigma = \mu_H + \mu_M \quad \text{and} \quad \tau = \frac{\mu_H}{\mu_H + \mu_M} \ .$$

The larger is σ for fixed τ, the better the discriminator.

The prior information for σ was gained from previous unrelated studies. It was deemed appropriate to use a uniform prior for σ for each word.

After examination of ninety words, Mosteller and Wallace thought it reasonable to assume that for writers of the same time period and similar style, as Hamilton and Madison were, the prior distribution of τ for any non-contextual word would be symmetric with much probability concentrated near 1/2. So τ was assigned a Beta distribution with equal parameters γ. Then γ was chosen of the form $\gamma = \beta_1 + \beta_2\sigma$ so that the variability of τ could be reduced with increasing σ. They concluded, after some empirical examinations, that a range for γ between 5 and 20 was plausible, and the analysis was carried out for several appropriate sets of β_1 and β_2. They later showed that the log odds vary up to 12% over this range, but never

enough to change the conclusions as to authorship of each essay.

Then the posterior distributions for (σ,τ) given that x_H and x_M are observed were found in the usual way: $p(\sigma,\tau|x_M,x_H) = C(x)p(\sigma,\tau)p(x_H,x_M|\sigma,\tau)$, where $C(x)$ is the appropriate constant. The mode of this density was used as the new estimate for σ and τ, and then new estimates for μ_H and μ_M could be calculated. The process tends to bring the estimate of τ closer to 1/2, thus making the estimates of μ_H and μ_M less discrepant.

Mosteller and Wallace anticipated many criticisms that might be aimed at their study. For example, they examined the validity of their assumption of independence of word rates, the plausibility of their prior distributions for σ and τ, and the change in an author's word-rate over time.

Their method assigned every disputed paper to Madison. When the procedure was applied to non-Federalist essays, it correctly classified both Hamilton's and Madison's, and gave ambiguous results for "neutral" papers, those written by neither man.

Although this is a very thorough study, it is not easily carried out. The number of computations required is so large that its use for resolving routine authorship disputes is not feasible.

2. Swinton and Whitman

In the nineteenth century, a new interest in linguistics developed, and one of the more popular studies was published by journalist William Swinton in *Putnam's Monthly Magazine* [6] in 1854. He later published a book, *Rambles Among Words*, [7], increasing his two original "rambles" to twelve.

This book would have probably been long forgotten if it were not for the fact that Swinton and Walt Whitman, also a journalist, were close friends during this period. Carroll Hollis [3] noted that this friendship began in 1855-56 and that Swinton became "Whitman's French tutor and translator, literary mentor, and philological expert, as well as Bohemian friend and family intimate."[3] Whitman had apparently always been interested in linguistics and had read extensively in this area. Hollis established, from examination of some original manuscripts, that his intensive philological investigations occurred in the early and middle months of 1856, when Swinton and Whitman were seeing each other often. This study seems to have been spurred on by Swinton's inspiration and enthusiasm. Some of the notes Whitman compiled became "An American Primer" [8], an unpublished collection dealing with the same kinds of topics discussed in *Rambles*. Other notes were collected in *Words*, a notebook which apparently served as a home-made filing system for his philological data.

---

[3] Hollis [2], p. 425.

According to Hollis, there is an unmistakable similarity in prose style, thought, and approach between "An American Primer" and in certain passages in *Rambles*. In addition, there appears a striking difference within *Rambles* itself between these passages and others. This difference, says Hollis, "...is so glaring that Swinton could have defended his sole authorship only on the basis of a split personality."[4]

Hollis also points out several minor items which indicate Whitman's assistance in the preparation of *Rambles*. For example, since Whitman's customary "Programme" and "Premonitary" replace "Table of Contents" and "Introduction," he feels that Whitman even saw to the printing. Corroboration of Hollis' speculation has been lent by other Whitman scholars who have independently examined *Rambles*.

Hollis' theory about this strange occurrence is as follows:

> "...Swinton had made arrangements to have an enlargement of his original articles printed as a book. Faced with dead-line pressures, he appealed to Whitman for help and got it via the notes made in the "Primer" and *Words* enterprises. Whether in deference to Whitman or because there was no time to do otherwise, he used these notes pretty much as he got them..."[5]

He also suggests possible explanations for Whitman's failure to get recognition for his contributions: a fear of endangering Swinton's position with the newspaper or school where he was employed, a desire on Whitman's part to repay Swinton for his encouragement and education,

---

4 *Ibid.*, p. 439.
5 *Ibid.*, p. 446.

or a refusal of the editor to acknowledge Whitman on account of notoriety which apparently was connected with his name at the time.

In this paper, an effort will be made to support, with some reservations, Hollis' hypothesis on the basis of statistical analyses of certain aspects of Swinton's and Whitman's styles, and that of the disputed passages appearing in *Rambles Among Words*.

3. The Analysis

The approach taken in the analysis was to amass the evidence from statistical tests of several characteristics of style with the hope that the "weight of evidence" would clearly point to one or the other of our authorship contestants for the questionable passages of *Rambles Among Words*. A control was needed in order to determine what actually was a Whitman- or Swinton-like style. Chosen for these purposes were the writings from each author most similar in topic and time of publication to that of *Rambles*. The essays used were the two previously mentioned *Putnam* articles [6] and another about novels [5] by Swinton and "An American Primer" by Whitman. (In fact, "Primer" should have extraordinarily close ties to *Rambles* if Hollis' supposition is correct, i.e., if "Primer" is actually the remainder of Whitman's notes on "words" after Swinton had taken part for use in *Rambles*). The Swinton sample contains about 7000 words, while the Whitman sample was somewhat larger, about 9000.

The statistical tests in this study with the exception of one, were developed with the advice of a student of Whitman's writing style.[6]

---

[6] Robert Ward Hutten, Department of English, University of North Carolina.

An attempt was made to quantify the "clues" which a Whitman scholar's intuition senses for authorship identification and then to perform statistical analyses on these data. Often this procedure is impossible. For example, Whitman's tendency to coin his own words is easy to recognize, but virtually impossible to test for statistically because of the low frequency of such words. Moreover, this approach sometimes leads to tests which do not discriminate well between the two authors in question. For example, the observation that Whitman tends to saturate his writing with nouns lead to a test of proportions of nouns used by Whitman and Swinton. In fact, however, the overall proportions of nouns used by the two men were quite similar. The difference lay in the fact that Whitman's noun-proportion in each sentence was erratic due to his frequent use of seriation. (This characteristic is discussed further in 3.3).

Caution should be exercised in order that the identification of distinctions between two authors' stylistic traits not be made solely from examination of the control group samples, since this is analogous to looking at the data before deciding what hypothesis to test. Rather, the knowledge about an author's style should be based on acquaintance with a variety of his works, which explains the need for the literary expert.

The four tests performed in this study are described below. The first is Yule's conventional sentence-length test. The remaining three capitalize on some notable aspects of Whitman's style, which is quite distinctive.

3.1 The sentence-length test was performed because the necessary data was readily available and because the idea of testing for a difference in two distributions is appealing to a statistician. The idea that a writer's sentence-length distribution remains the same throughout all his writing is most likely false. In fact, the idea that a writer's sentence-length distribution remains the same throughout his writing on the same topic is questionable. (See 4.1 for further discussion on this point.) However, with this hope, a Kolmogorov-Smirnov two-sample test was used in the comparison of sentence-length distributions.

First the distributions of the Whitman and Swinton samples were compared with each other to see if there actually does appear a difference on this variable between the two men's styles. The next step was to compare each control sample with *Rambles*.

One problem arising here is the assumption required for the Kolmogorov-Smirnov test that the observations in each sample are independent. This assumption is obviously not true in the sentence-length case. For example, an author may group short sentences for emphasis. ("The words of the body! The words of parentage! The words of husband and wife! The words of offspring! The word Mother! The word Father!"[7]). For this reason, paragraphs (such as Whitman's above) which contain series of sentences which are definitely not independent were excluded from the samples.

---

[7] Whitman [8], p. 571.

3.2 The second test used in the analysis was developed to check the disputed passage for an unusual characteristic of Swinton's style. Swinton writes in long, complex sentences which are often begun with a connecting word or phrase (e.g., then, thus, on the other hand). Whitman, however, tends to write more consistently with the conventional sentence-opening noun. So an effort was made to quantify this characteristic so that it could be tested for in the *Rambles* data.

A Bernoulli random variable was associated with the first word of each sentence. This random variable takes on the value 1 if the word is an article, noun, or a subjective pronoun (category A), a 0 otherwise (category B). The sum of n independent Bernoulli random variables has a binomial (n,p) distribution where p = P[category A word occurs]. The random variables connected with consecutive sentences are not likely to be independent since the structure of a sentence often depends on the structure of the one preceding it. Therefore, the Bernoulli variables were placed in random order before being added in groups of about thirty to form Bernoulli random variables. The normal approximation to the binomial was used, and these normal samples were tested for differences in means by a t-test. The problems of appropriateness of the normal approximation and the assumption of equality of variances will be discussed in 4.2.

3.3 The third test seeks to quantify Whitman's unusual tendency to use many nouns, often strung together in series. ("Words follow character-maturity, independence, individuality." "Why are names so mighty? - because facts, ancestry, maternity, faiths, are."[8])

---

[8] *Ibid.*, p. 570.

This test is similar to test 1 in that distributions of random variables were compared. The random variable used in this case was the proportion of nouns in each sentence. Again, a Kolmogorov-Smirnov two-sample test was employed.

The same problem with the independence assumption as was mentioned in 3.1 arises in this case. A different approach for solution of this problem is to choose a sample of observations which are not affected by each other from the data; i.e., which are, in a non-technical sense, independent. One method to approximate this desired effect would be to choose a relatively small simple random sample from the original data. The disadvantage here is that much data is lost.

3.4 The fourth test was developed in an effort to compare quantitatively Whitman's smooth, simple style with Swinton's complex style. Swinton's writing is full of parenthetical expressions and phrases which are set off by commas, making his writing sound heavy and pedantic. ("Thus, in literature, every longing and susceptibility of the soul, and, in fact, every mental want, creates for itself a satisfaction and supply."[9])

For this comparison, all punctuation marks except end punctuation (i.e., commas, semicolons, colons, dashes) were classified into one of two categories. Category A was "connective" punctuation, those marks of punctuation which stand in place of, or along with, a conjunction. Category B, then, consisted of punctuation used to set off clauses, phrases, or a single word for non-connective purposes.

---

[8] Swinton, [5], p. 339.

Again, as in test 2, each punctuation mark was associated with a Bernoulli random variable X such that

$$X_i = \begin{cases} 1 & \text{if i-th punctuation mark } \epsilon \text{ Category A} \\ 0 & \text{otherwise .} \end{cases}$$

Then the analysis was carried out, with minor changes, as in test 2.

4. Results

In this section, the results of the four tests described above are reported.

4.1 In order to check the question as to whether a writer's sentence-length distribution remains the same throughout his writing on the same topic, the data from the three Swinton essays were tested against each other in pairs by a Kolmogorov-Smirnov two-sample test. Sample S1, ("Novels") consists of 111 observations, Sample S2 (*Putnam* essay 1) has 40, and Sample S3 (*Putnam* essay 2) has 50. For example, to test

$$H_0: F_{S1}(x) = F_{S2}(x) \qquad H_1: F_{S1}(x) \neq F_{S2}(x) \qquad \alpha = .05 .$$

The test statistic $D_{S1,S2} = .233 < .251 = D_{.95;111,40}$. Therefore, H_0 cannot be rejected. The same holds true for the other two pairs, giving results of $D_{S2,S3} = .200 < .288 = D_{.95;40,50}$ for samples 2 and 3, and $D_{S1,S3} = .099 < .232 = D_{.95;111,50}$ for 1 and 3. So in this case at least, the writer appears consistent in his sentence-length.

The combined Swinton sample contains 201 observations, Whitman's has 261, and *Rambles* has 322. Histograms for the three sets of data are shown in Figures 1 - 3.

First the Swinton and Whitman samples were compared:

$$H_0: F_S(x) = F_W(x) \qquad H_1: F_S(x) \neq F_W(x) \qquad \alpha = .05$$

$D = .300 > .128 = D_{.95;261,201}$. H_0 is rejected; there is a distinction in the sentence-length distributions of the two men.

Next the Swinton and Whitman data are each tested against the *Rambles* sample. The results are: (1) Swinton-*Rambles*; $D = .296 > .122 = D_{.95;201,322}$ (2) Whitman-*Rambles*; $D = .066 < .113 = D_{.95;261,322}$. That is, the hypothesis of equality of the Swinton and *Rambles* distributions is rejected, while the hypothesis of equality of the Whitman and *Rambles* distribution is not.

The comparison of log-sentence-length and the normal distribution, which was suggested by Williams, was also carried out. Figure 4 is a histogram of the log-sentence-length data for the Whitman sample. Contrary to Williams' opinion, neither Whitman's nor Swinton's sample was found to be normal as tested by a Kolmogorov-Smirnov-Lilliefors test. For the Swinton sample, the test statistic $D = .102 > .062 = D_{.95;201}$. For the Whitman sample, $D = .066 > .052 = D_{.95;261}$. A square-root transformation yielded smaller, but still significant values for $D(D_{Swin} = .070, D_{Whit} = .088)$. Figure 5 is the histogram for square root-sentence-length for the Whitman data.

4.2 The random variable associated with the first word of every sentence was a Bernoulli random variable

$$X_i = \begin{cases} 1 & \text{if the first word of sentence } i \in \text{Category A} \\ 0 & \text{otherwise .} \end{cases}$$

The 261 observations from Whitman, 201 from Swinton, and 322 from *Rambles* were each arranged in random order and then added in groups of about 30. Since the sample values of $p = Pr[X=1]$ was always between 1/3 and 2/3, the normal approximation is a safe one. Let $Y = \sum_{i=1}^{n} X_i$. Then $Y/n \sim N(p,pq/n)$.

The three sets of data are shown below:

Sample 1 (Swinton)

$Z_S = Y/n$
.367
.500
.290
.414
.300
.320
.346

$N_S \equiv$ Number of observations in Swinton sample = 8

$$\bar{Z}_S = \frac{\sum_{j=1}^{N_S} Z_{S_j}}{N_S} = .362$$

$$N_S S_S^2 = \sum_{j=1}^{N_S} (Z_{S_j} - \bar{Z}_S)^2 = .0336$$

Sample 2 (Whitman)

$Z_W = Y/n$
.567
.633
.533
.567
.667
.700
.733
.500
.633
.538

$N_W = 10$

$$\bar{Z}_W = \frac{\sum_{j=1}^{N_W} A_{W_j}}{N_W} = .607$$

$$N_W S_W^2 = \sum_{j=1}^{N_W} (Z_{W_j} - \bar{Z}W)^2 = .0555$$

Sample 3 (*Rambles*)

$Z_R = Y/n$	$N_R = 11$
.467	
.500	$\bar{Z}_R = \frac{\sum_{j=1}^{N_R} Z_{R_j}}{N_R} = .464$
.500	
.433	$N_R S_R^2 = \sum_{j=1}^{N_R} (Z_{R_j} - \bar{Z}_R)^2 = .0325$
.400	
.533	
.533	
.367	
.433	
.533	
.400	

The several normal observations from each sample were tested in various pairs for equality of means by a t-test. The assumption of equality of variances is not badly violated if p is in the "middle" range as it appears to be. For example, if $1/3 < p < 2/3$, then $.086 < \sigma < .091$, a difference of only .005.

First, the test was performed on the Swinton and Whitman data to see if there seemed to be a difference between the two.

$H_0: p_1 = p_2$ $\qquad$ $H_1: p_1 \neq p_2$ $\qquad$ $\alpha = .05$

where $X_W \overset{d}{=}$ Bernoulli (p_1) and $X_S \overset{d}{=}$ Bernoulli (p_2)

Then

$$\frac{(\bar{Z}_W - \bar{Z}_S)\sqrt{N_W + N_S - 2}}{\sqrt{\frac{1}{N_W} + \frac{1}{N_S}}\ \sqrt{\sum (Z_{W_i} - \bar{Z}_W)^2 + \sum (Z_{S_i} - \bar{Z}_S)^2}} \sim t_{N_W + N_S - 2}\ .$$

Here $|t| = |-6.92| > 2.12 = t_{.95;16}$. Therefore H_0 is rejected; i.e., there does appear a difference between Swinton and Whitman on this characteristic.

Next the two authorship candidates were tested against the *Rambles* passage. The Swinton-*Rambles* t-test strongly rejected the hypothesis of equality of means with the test statistic $|t| = |-3.54| > 2.11 = t_{.95;17}$. But the Whitman-*Rambles* data even more strongly rejected the hypothesis of equality with $|t| = |4.83| > 2.09 = t_{.95;19}$. Therefore, the conclusion must be that the *Rambles* passage is not strikingly similar to either author.

4.3 Since each sentence served as an observation for the third test, there were 201 observations in the Swinton sample, 261 observations in the Whitman sample, and 322 observations in the *Rambles* sample. The histograms for this data may be found in Figures 6-8.

The first test here, as in 4.1, was to run the Kolmogorov-Smirnov analysis on pairs of the three Swinton essays to see if they were similar in the noun usage characteristic. For example,

$$H_0: F_{S1}(x) = F_{S2}(x) \qquad H_1: F_{S1}(x) \neq F_{S2}(x) \qquad \alpha = .05 .$$

Since $D_{S1,S2} = .192 < .251 = D_{.95;111,40}$, H_0 is accepted. The same is true for the other comparisons, since $D_{S1,S3} = .217 < .232 = D_{.95;111,50}$ and $D_{S2,S3} = .095 < .288 = D_{.95;40,50}$. So apparently Swinton is consistent among essays with regard to this characteristic.

The next step was to test for a difference between Swinton's and Whitman's noun usage. Since $D = .230 > .128 = D_{.95;201,261}$, the

hypothesis of equality of the two distributions must be rejected; i.e., Swinton and Whitman appear to differ on this characteristic.

When the *Rambles* passage was tested against Swinton's writing, the two were found to be significantly different since $D = .244 > .122 = D_{.95;322,201}$. However, the hypothesis of equality of the distributions of *Rambles* and Whitman cannot be rejected since $D = .071 < .113 = D_{.95;322,261}$. So the *Rambles* passage appears to be more similar to Whitman's writing than to Swinton's on the characteristic of the proportion of nouns per sentence.

4.4 The analysis was carried out here as in test 2, but, this time, with the category of every punctuation mark being an observation of a Bernoulli random variable, X_i. The 723 observations from Swinton, 750 observations from Whitman, and 528 from *Rambles* were first put in random order. They were then added in groups of about 30 for *Rambles* and Whitman and about 50 for Swinton. The larger number was used in the Swinton case to improve the normal approximation since the Category A observations seemed sparse. Let $Y = \sum_{i=1}^{n} X_i$. Then $Y/n \sim N(p,pq/n)$.

The three sets of data are shown below:

Sample 1 (Swinton)

$Z_S = Y/n$

"Novels"	*Putnam's* "Rambles"
.26	.34
.18	.24
.26	.20
.36	.34
.29	.28
.48	.49
.30	.32
	.21

$N_S = 15$

$$\bar{Z}_S = \frac{\sum_{j=1}^{N_S} Z_{S_j}}{N_S} = .30$$

$$N_S S_S^2 = \sum_{j=1}^{N_S} (Z_{S_j} - \bar{Z}_S)^2 = .150$$

Sample 2 (Whitman)

$Z_W = Y/n$	
.59	.66
.70	.45
.54	.63
.58	.53
.50	.57
.67	.57
.57	.66
.63	.54
.51	.59
.43	.59
.50	.51
.57	

$N_W = 23$

$$\bar{Z}_W = \frac{\sum_{j=1}^{N_W} Z_{W_j}}{N_W} = .57$$

$$N_W S_W^2 = \sum_{j=1}^{N_W} (Z_{W_j} - \bar{Z}_W)^2 = .087$$

Sample 3 (*Rambles*)

$Z_R = Y/n$	
.37	.53
.43	.63
.40	.53
.57	.40
.43	.43
.57	.47
.40	.46
.43	.50
.50	.65

$N_R = 18$

$$\bar{Z}_R = \frac{\sum_{j=1}^{N_R} Z_{R_j}}{N_R} = .48$$

$$NS_R^2 = \sum_{j=1}^{N_R} (Z_{R_j} - \bar{Z}_R)^2 = .123$$

Again, a t-test was used to test equality of means, and again the violation of the assumption of equality of variances should make little difference, since the most extreme sample value of $p = P[X_i = 1]$ is .30.

First, Swinton's "Novels" and *Putnam's* "Rambles" observations were tested against each other to see if they were consistent with regard to this characteristic. The test statistic $t = 0.54 < t_{.95;13}$, failing to reject the hypothesis of equality of means at the .05 level.

But, as in test 2, both the Whitman-*Rambles* ($t_{W,R} = 3.89 > 2.33 = t_{.95;39}$) and Swinton-*Rambles* ($t_{S,R} = 5.48 > 2.35 = t_{.95;31}$) hypotheses of equality were rejected.

In another approach to analysis of the same data, a χ^2-test was employed, giving the same results. That is,

"Novels" - *Putnam* "Rambles"	$\chi^2 = .025 < 3.84 = \chi^2_{.95;1}$
Whitman-*Rambles*	$\chi^2 = 18.94 > 3.84 = \chi^2_{.95;1}$
Swinton-*Rambles*	$\chi^2 = 40.51 > 3.84 = \chi^2_{.95;1}$

5. Conclusions

The results seem to contradict Hollis' supposition that the *Rambles* passages were lifted, with no changes, from Whitman's notes. The passages are too different from "An American Primer" in certain characteristics, as indicated by tests 2 and 4, to simply be two parts of a whole piece. The differences which showed up between *Rambles* and the Swinton essays on tests 2 and 4 should be more expected since they weren't claimed to be part of the same article.

However, the results of tests 1 and 3 do support the theory that there is a similarity in style between *Rambles* and Whitman's writing. This support is strengthened by the observation that the characteristics

appear to be consistent within one writer's style, as indicated by the tests of "Novels" against the other Swinton *Putnam* articles.

So the similarities of *Rambles* to Whitman appear to be more than coincidental, yet the two are not identical. There are several possible explanations. One is that Swinton used Whitman's notes heavily, but rewrote certain parts of them, producing a blend of their two styles. Another explanation is that Hollis' list of Whitman passages in *Rambles Among Words*, which was used for the *Rambles* sample, includes some mistakes, producing results falling halfway between the styles of the two men on some statistical tests.

A statistical procedure cannot, of course, produce an unquestionable answer to a problem such as this one. Since results of the four tests in this case are not unanimously supportive of one or the other side, the interpretation is even more difficult and subject to individual judgement. However, the statistical procedures used do provide useful evidence, which considerably assists our assessment of the situation.

BIBLIOGRAPHY

1. Brinegar, Claude (1963). "Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship," *JASA*, Vol. 58, 85-96.

2. Conover, W.J. (1971). *Practical Nonparametric Statistics*, New York, John Wiley and Sons, Inc.

3. Hollis, Carroll (1959). "Whitman and William Swinton: A Co-operative Friendship," *American Literature*, Vol. 30, 425-449.

4. Mosteller, Frederick and Wallace, David (1964). *Inference and Disputed Authorship: The Federalist*, Reading, Mass., Addison-Wesley Publishing Company.

5. Swinton, William (1954). "Novels: Their Meaning and Mission," *Putnam's Magazine*, Vol. 4, 389-396.

6. Swinton, William (1854). "Rambles Over the Realms of Verbs and Substantives," *Putnam's Magazine*, Vol. 4, 472-481, 602-609.

7. Swinton, William (1859). *Rambles Among Words*, New York, Scribner.

8. Whitman, Walt (1855?) "An American Primer," ed. Horace Traubel, *The Poetry and Prose of Walt Whitman*, 569-585.

9. Williams, C.B. (1956). "Studies in the History of Probability and Statistics," *Biometrika*, Vol. 43, 248-256.

10. Williams, C.B. (1939). "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style," *Biometrika*, Vol. 31, 356-361.

11. Yule, G. Udny (1938). "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship," *Biometrika*, Vol. 30, 363-390.

12. Yule, G. Udny (1944). *The Statistical Study of Literary Vocabulary*, Cambridge University Press.

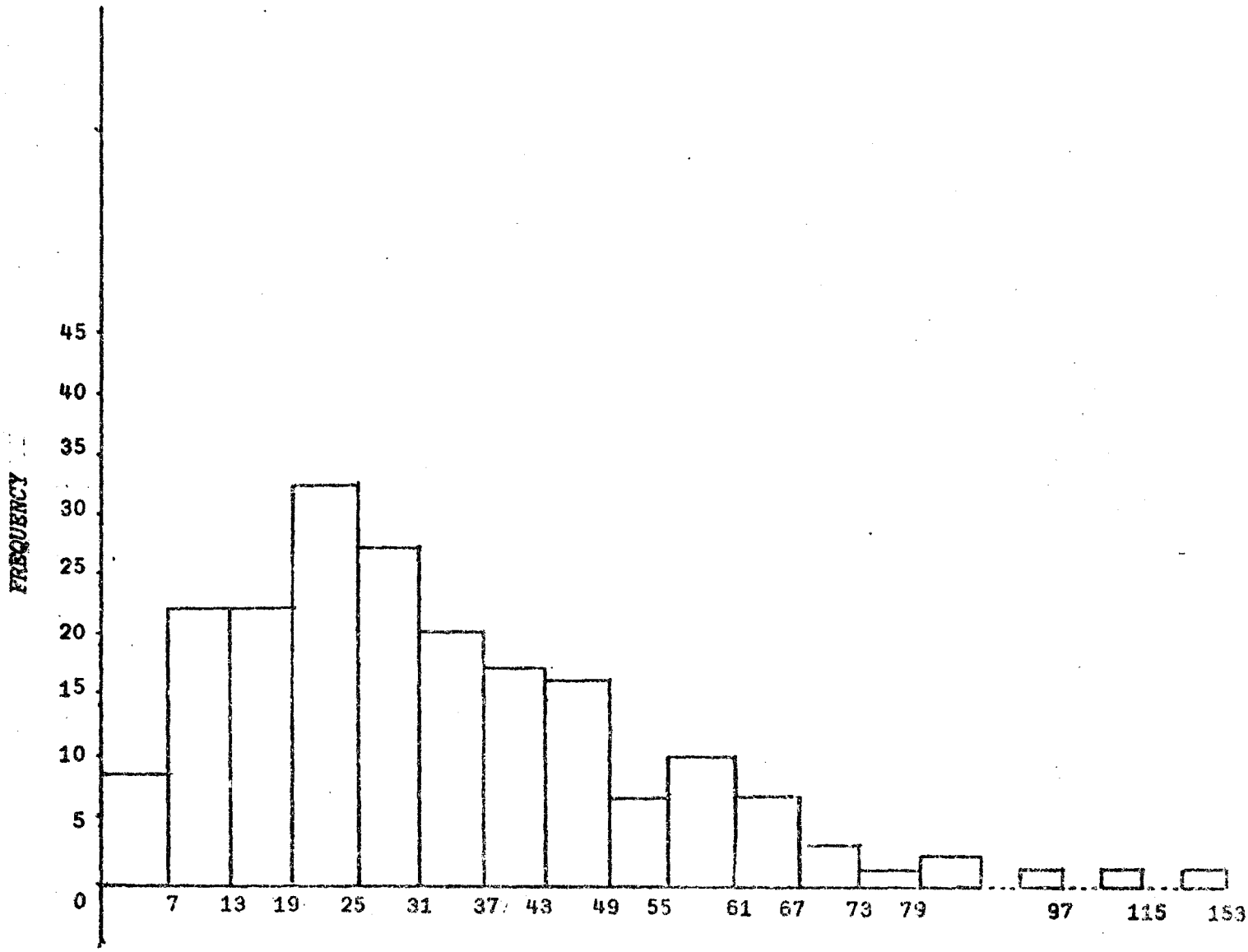

Figure 1

Swinton sentence-length

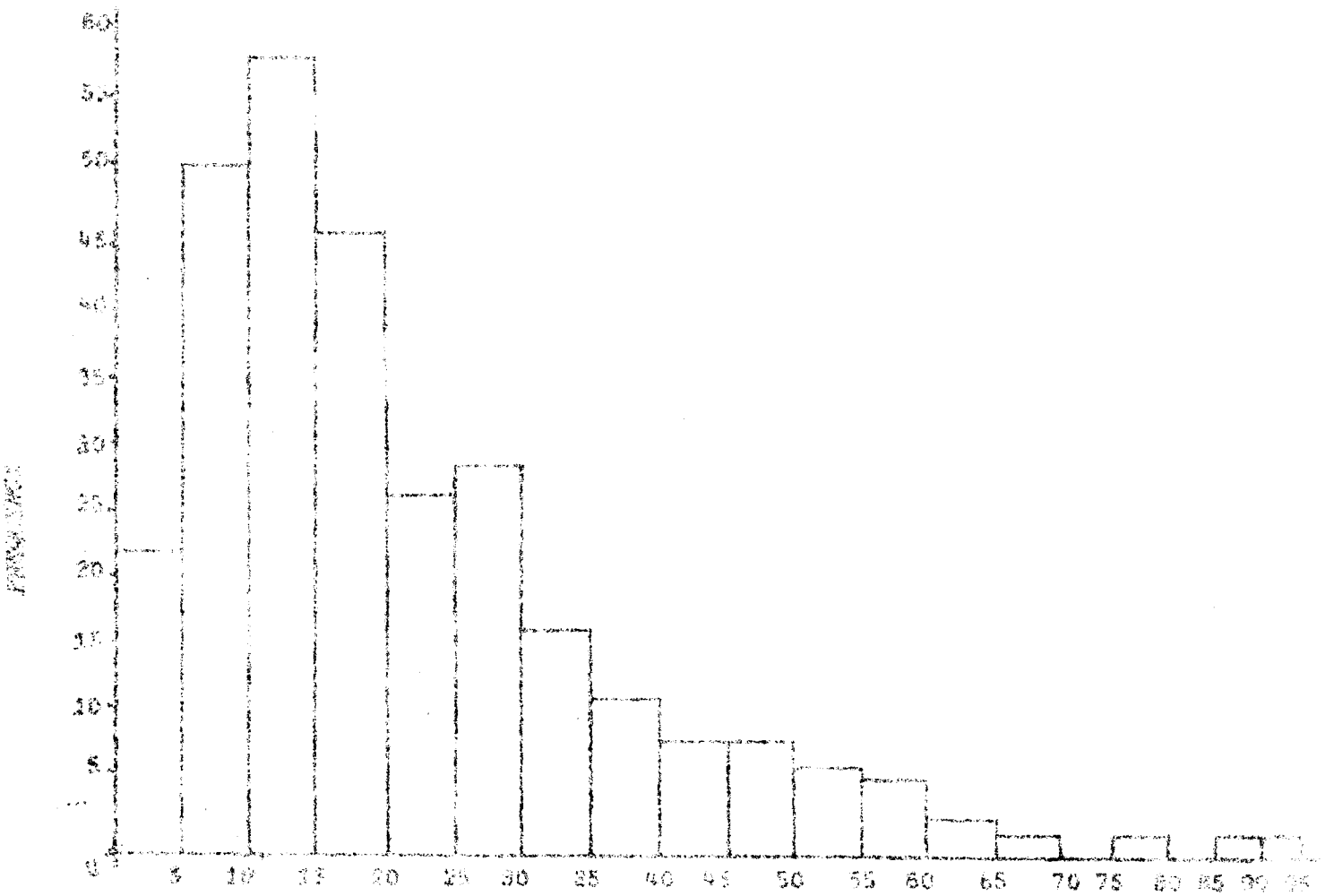

Figure 7

Whitman sentence-length

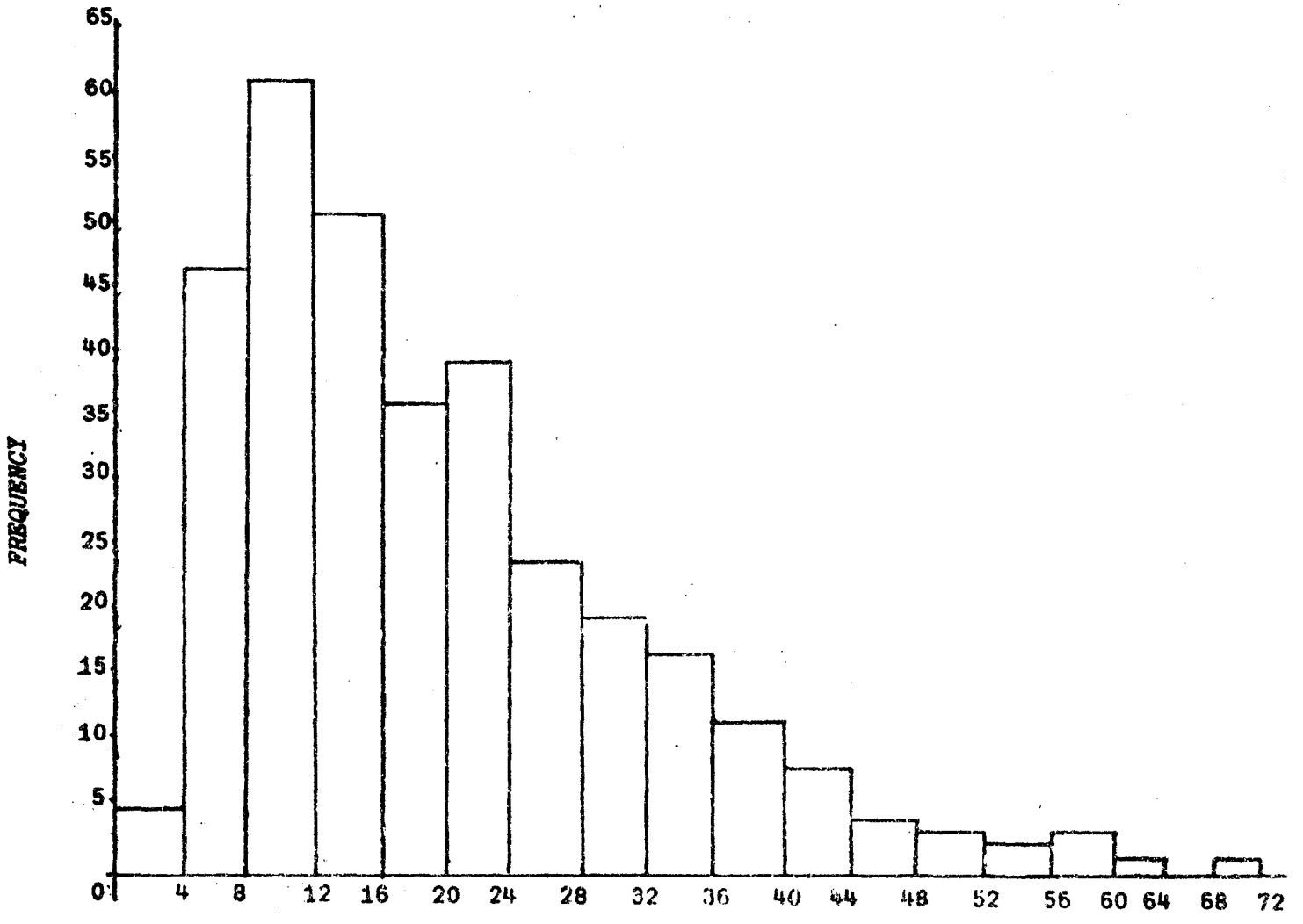

Figure 3

Rambles sentence-length

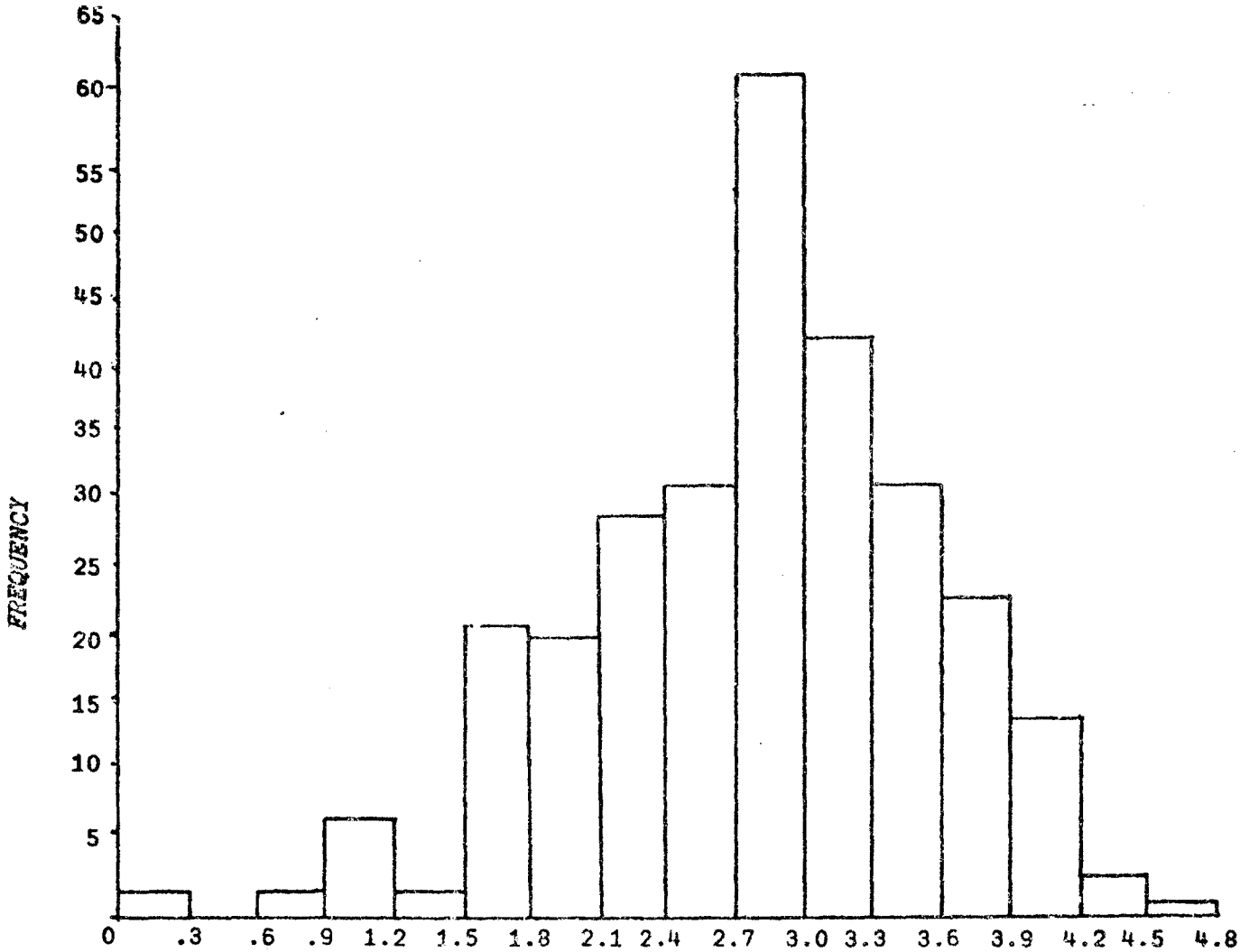

Figure 4

Whitman Log sentence-length

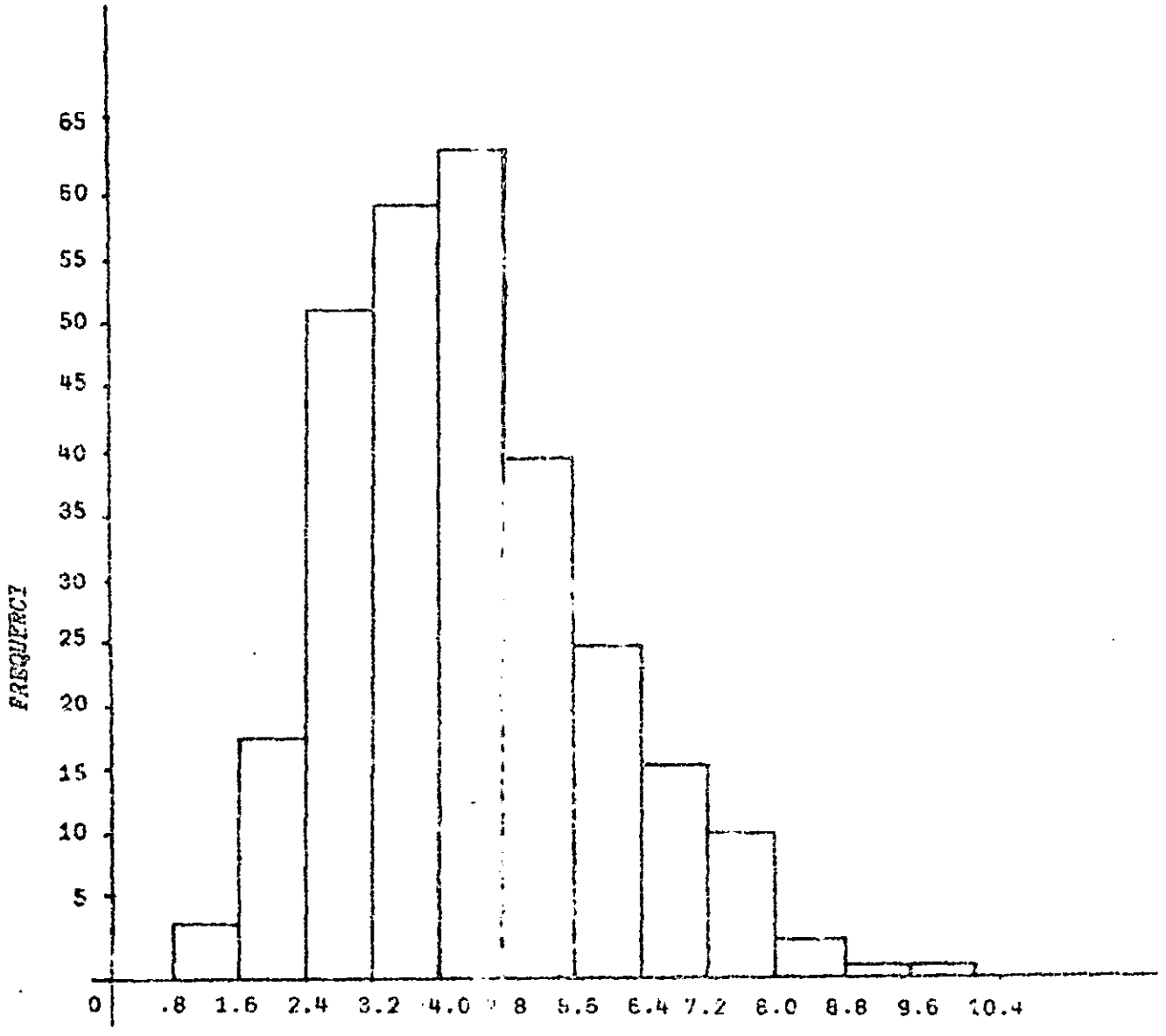

Figure 5

Whitman Square root-sentence-length

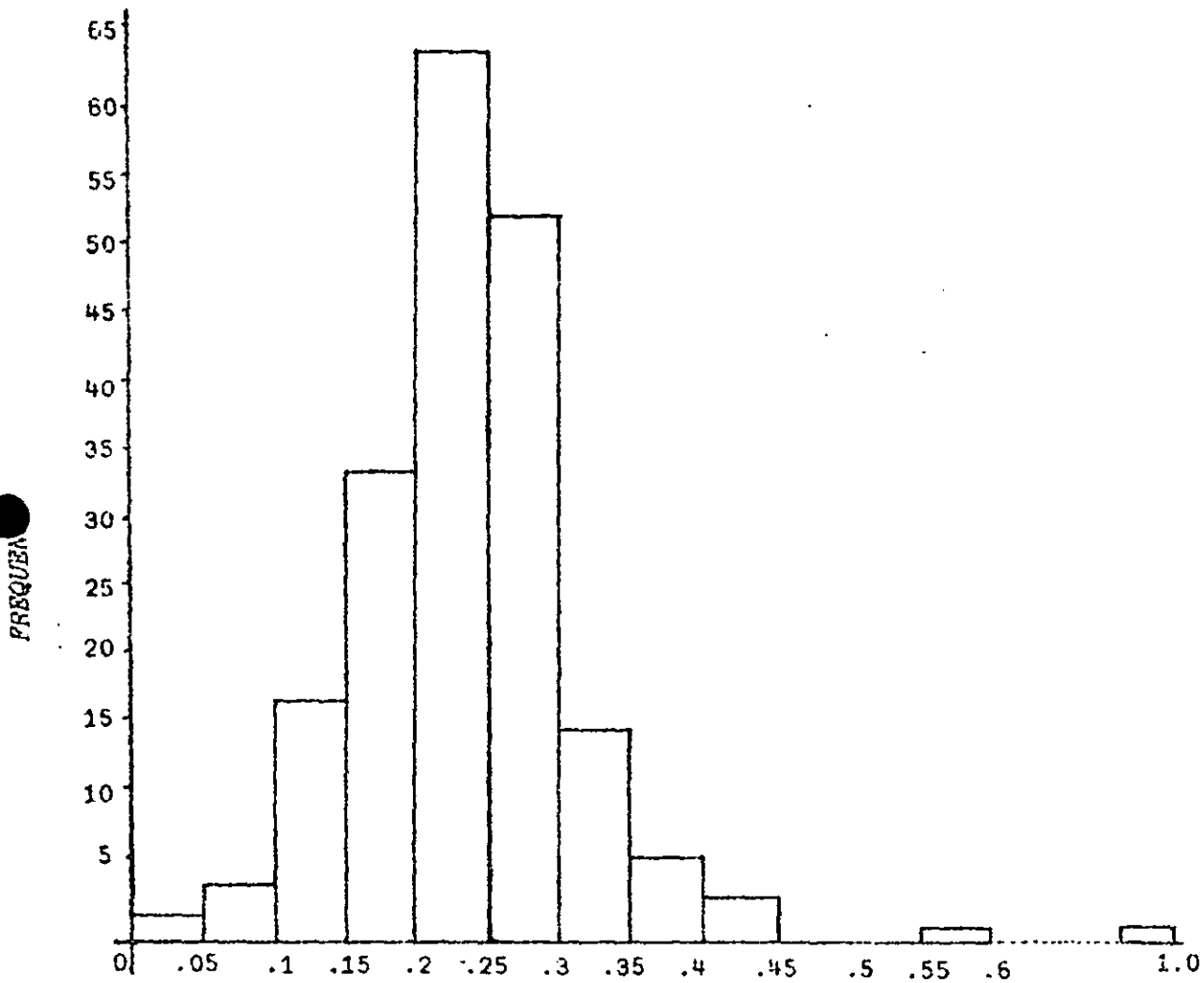

Figure 6

Swinton noun proportion per sentence

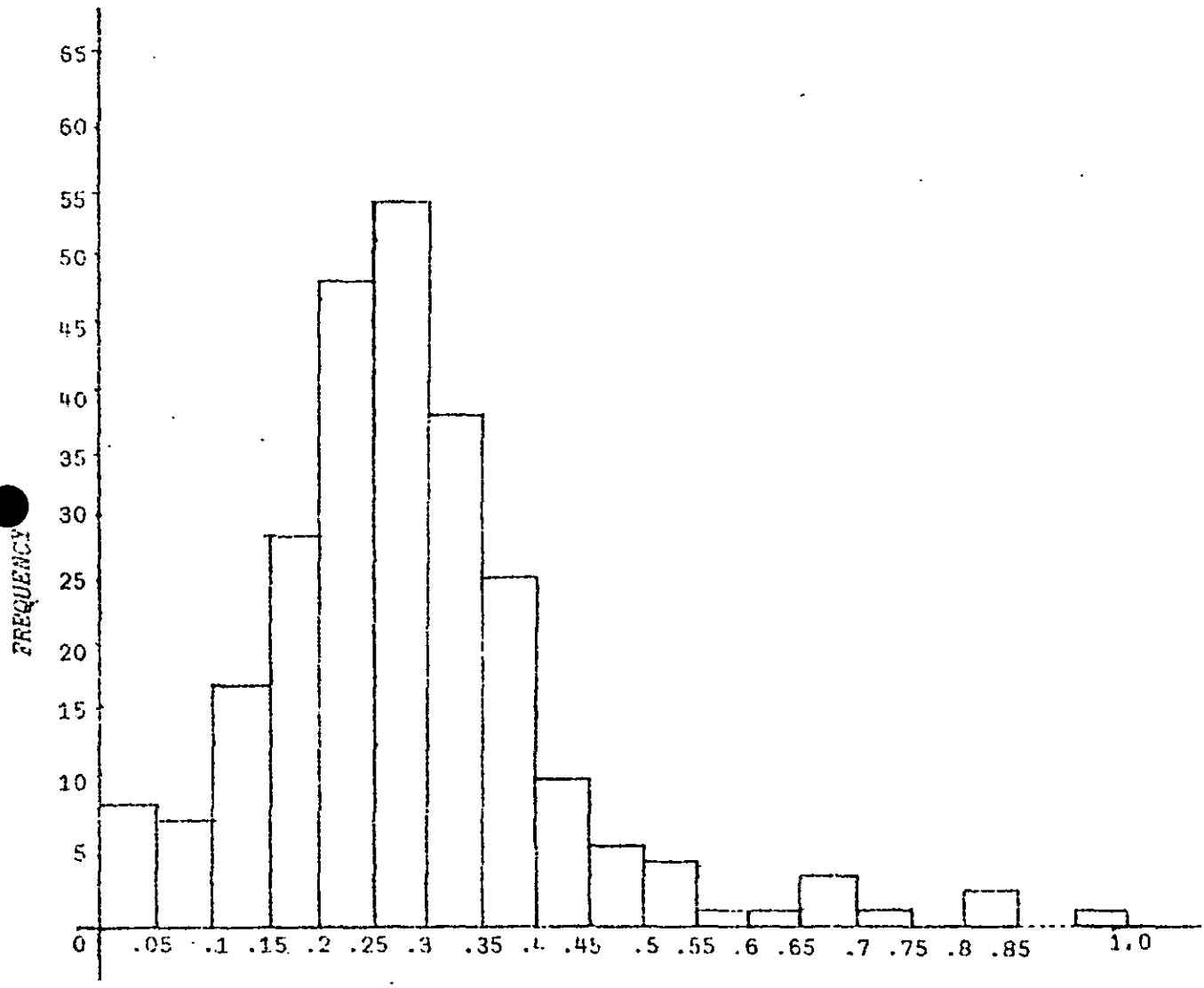

Figure 7

Whitman noun proportion per sentence

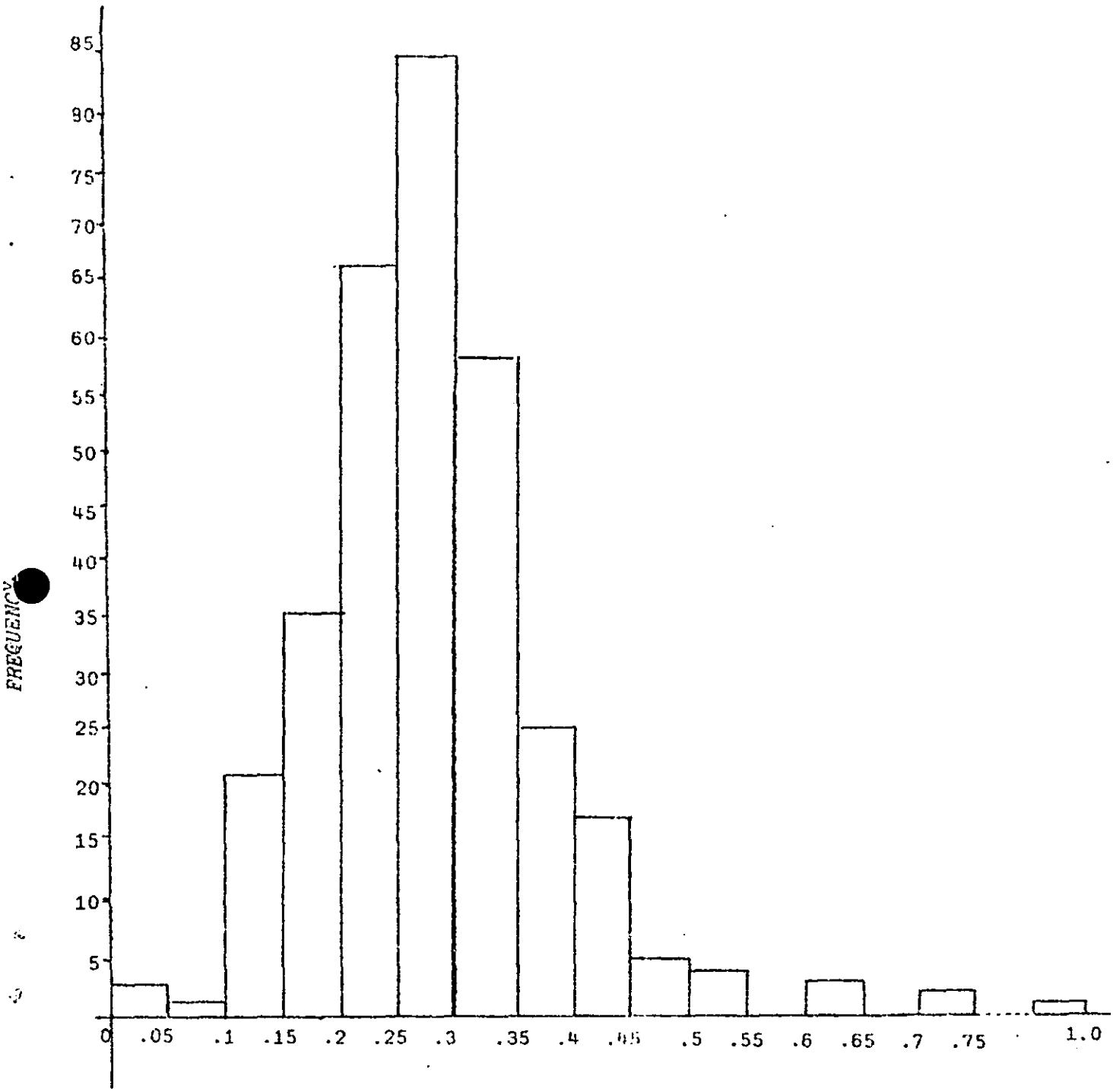

Figure 8

Rambles noun proportion per sentence