Mackey, Walter H., II

A MONTE-CARLO INVESTIGATION OF STATISTICAL POWER UNDER THE MIXED-NORMAL DISTRIBUTION

Wayne State University PH.D. 1982

A MONTE-CARLO INVESTIGATION OF STATISTICAL POWER UNDER THE MIXED-NORMAL DISTRIBUTION

by

Walter H. Mackey II

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

1982

MAJOR: EDUCATIONAL EVALUATION AND RESEARCH

Approved by:

Joseph Z Ponessa Jr 6/22/82
Adviser date

Donald C Manoukian

Thomas J Lyons

Eugene P Smith

ACKNOWLEGEMENTS

Undoubtedly, the work required for this dissertation could never have been completed were it not for the time and efforts of several other people. I wish to give special thanks to the following:

To Dr. Joseph Posch, Jr., my major adviser, whose direction and personal attention during the final stages of my doctoral program made this a reality;

To Dr. Donald Marcotte, Dr. Eugene P. Smith, and Dr. Thomas Duggan, the other members of my committee, whose many insights and comments provided immeasurable help in my work;

To Mr. Larry Fromm of the University of Michigan, whose expert assistance aided in preparing of the computer simulations;

To Mrs. Karen Leigh, whose typing and clerical skills put the preliminary and final drafts of this manuscript in such beautiful form;

To the many people from Wayne State University, the University of Michigan, the University of Detroit, and Ohio State University who enabled me to overcome many obstacles, both great and small.

To Dr. Maureen Sie, whose encouragement in the beginnings of my doctoral studies gave me the confidence to continue.

A final special thanks must go to my mother, Mrs. Lucille Mackey. If I attempted to list the reasons, these pages would never finish.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

STATEMENT OF THE PROBLEM UNDER INVESTIGATION

Introduction

Truth is not static. Absolutes are becoming fewer as technology and instant communication are causing society to become less willing to posit definite conclusions. Results are often prefaced with conditions, and thoughts that are the reality of today may well be the folklore of tomorrow.

Relativism is most evident in the realms of science, where sacrosanct conclusions and theories are constantly being revised as the quest for truth leads to new and previously unattainable horizons. The field of statistics shares these same uncertainties. The development of sophisticated computer technologies has enabled practitioners and theorists of the statistical sciences to attempt investigations of questions that were, for all practical purposes, unresearchable due to the restraints of time and resources. The present work will look at one such question and give direction for possible future research.

Generalized Purpose of the Study

The present study is a comparison of the relative effectiveness of two common statistical tests. One, the "Student" *t*-test, is based upon an application of a normal curve. Its distribution-free counterpart, the Mann-Whitney *U*-Test, does not require the use of a theoretical distribution in order for a researcher to make a decision about a hypothesis under investigation. Both of these tests are used to detect significant differences between two population means based on the observed difference between two samples. The method of analysis used in both the *t* and *U*-tests is commonly used to draw inferences about Experimental-Control group studies.

The model of a normal population will be adjusted to varying degrees, and the ability of these two tests to correctly identify significant differences will be estimated by a computer simulation approach. A specific family of curves, classified as mixed-normal, will serve as the population frequency distributions from which the samples are randomly drawn. The particular population shape that will be investigated is formed by composition of two separate curves, each of which is individually a form of the normal distribution.

Mathematical Models

Mathematicians build models. They develop their logical hypotheses from necessary assumptions that may, to varying degrees, be applicable to real life situations of research.

The gap between the applied and theoretical branches of knowledge is quite evident in statistics. The normal curve, so common in statistical models, is an abstraction. No real data sets are "normal" since normality requires a continuous and infinite distribution. The assumption of a normal curve is the most common base upon which many modern statistical tools have been developed. The general appropriateness of this model has been questioned ever since its discovery, even by its very creators. As a result, certain statisticians began to direct their energies toward what is commonly referred to as "nonparametric" statistics.

Classes of Inferential Statistics

The two terms, nonparametric and distribution-free statistics, are used by many in a synonymous fashion. However, this is technically incorrect. As Bradley points out, a nonparametric test is one that fails to make a hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test depends upon no precise assumptions concerning the form of the sampled population.[1] The Sign Test, for example, is a test that is distribution-free but at the same time cannot be classified as nonparametric. Zuwaylif offers the following distinction between parametric and nonparametric statistical tests: parametric tests are those in which the hypotheses deal with

---

[1] James V. Bradley, Distribution-Free Statistical Tests (Englewood Cliffs, New Jersey: Prentice-Hall, 1968), p. 15.

population parameters, whereas nonparametric tests deal with hypotheses concerning population frequency distributions.[1]

Historical Perspective

It is assumed by many that the distribution-free tests are recent developments, originating over the last fifty years. Actually, distribution-free tests trace their development back to the very beginnings of probability theory and combinatorial mathematics.

The history of statistics has developed through many channels with its origins being found in the early Renaissance era and the emergence of commercial insurance against risks in certain Italian cities. It was the "geometry of the die," the alea geometria, of Blaise Pascal, Pierre de Fermat, and Christian Huygens, which was the catalyst for further probability theory. The aristocracy and their interest in gaming theory and the laws of chance provided the impetus for great mathematicians as they began to formulate the body of knowledge that grew into formal probability and statistical modeling.[2]

The year 1733 could be cited as the birthdate of what was to be known as the normal curve.[3] At this time, deMoivre found the limiting form of a binomial distribution when the

---

[1] Fadil H. Zuwaylif, Applied Business Statistics (Reading, Massachusetts: Addison-Wesley, 1974), pp. 24-379.

[2] Bradley, pp. 1-2.

[3] Bradley, pp. 2-3.

number of trials became infinite. Further work by Gauss and LaPlace extended the concepts of normality to distributions of errors in astronomical observations.

Mathematically, these errors are equally likely to be positive or negative, and the frequency of these errors decreases as their number increases in absolute magnitude. The slope of the frequency distribution, dy/dx, is zero when the number of errors, x, equals zero. That is to say, the frequency distribution has a relative maximum at zero, the peak of the normal curve. Its slope is also zero when the frequency of errors, y, is zero as x approaches both positive and negative infinity. So the curve $dy/dx = -Cxy$ fulfills these conditions as the derivative, dy/dx, is equal to zero whenever x or y is zero. Since $dy/dx = -Cxy$, it follows from elementary calculus that $dy/y = -Cx\,dx$. Integrating each side of the equation gives $\ln y = (-Cx^2/2)+K$ or an equivalent form of $y = K\,e^{-(Cx^2/2)}$ which is essentially the equation of the normal curve.[1]

The functional formula for normality provides the foundation of parametric statistics. It is based on a precise and exact mathematical model derived from a set of rigid assumptions.

The normal distribution was found to closely approximate certain sets of observed data from several diverse fields of endeavor.

---

[1] Bradley, p. 3.

During the period preceding the French Revolution, the leaders of thought in the Western World held a deterministic view of the universe.[1] God was viewed as the great Mathematician-Architect, and it was felt that society's eventual quest was to uncover those great laws of order and truth which dictated the course of human events. It was under these philosophical principles that the works of earlier statisticians such as LaPlace dominated throughout the nineteenth century.

The normal curve was considered to have almost divine properties, and the mean and standard deviation became the generally accepted standards of central tendency and variability. The mathematical model of normal curve was accepted as the standard to which many distributions were compared.

Bradley refers to the frequent application of normality as the "Gaussian Grip of the Normal Mystique," the attempt to correctly or incorrectly fit many empirical distributions to an absolute standard.[2] Nothing is "normal" in the mathematical sense of the word since no real data takes the shape of an infinite distribution. Neither is the mean the best measure of "average", especially for data that is highly skewed. The standard deviation and variance have precise mathematical properties when related to the normal model, but there is no guarantee that these measures of dispersion possess attributes

---

[1] "Mathematical Probability," Encyclopedia Brittanica, 18, 1970 ed., pp. 570-79.

[2] Bradley, pp. 13-4.

that preclude the use of such concepts as average deviation or range for certain data sets.

The communication gap between mathematical theorist and social scientist practitioner began to widen. The normal curve and its adjuncts were applied without much insight in a feverishly "cookbook" approach as more empirical distributions were found to be non-normal upon careful analyses. This apparent deviation from normality did not deter the further development of what are generally classified as parametric tests. In time, the belief in absolute normality was somewhat tempered. Bradley has labeled this stance as "quasi-universal quasi-normality."[1] However, as the twentieth century emerged, the classical parametric tests of Fisher, Pearson, and Gossett were originally developed using the assumption of sampling from a normally distributed population of infinite range.

Distribution-Free Statistics

Savage cites the formal beginning of the field of nonparametric statistics as 1936, although certain methods of distribution-free analysis are as old as formal probability theory itself.[2] Savage considers 1936 as the "true beginning

---

[1] Bradley, p. 6.

[2] I. Richard Savage, "Bibliography of Nonparametric Statistics and Related Topics," American Statistical Association Journal, 48 (1953), pp. 844-906.

of the subject". It was at this time that Hotelling and Pabst published their paper on rank correlations.[1]

The late nineteen forties and nineteen fifties saw a great increase in both the development and popularity of non-parametric tests. Several testing procedures, quite sophisticated in nature, were developed. Instead of integrating over mathematical density function in order to determine a probability, statisticians were now developing tests that could test hypotheses by simple probability statements that were derived from a finite number of combinatorial arrangements. Several textbooks of varying degrees of complexity were authored, most of which came out strongly in favor of the use of these distribution-free tests in the educational, behavioral, and social sciences. This advocacy of distribution-free analysis was especially true when some of the rather stringent parametric assumptions were overtly violated or even somewhat in doubt.

Many practitioners were concerned that the trend toward distribution-free statistics was forcing them to discard important "information" when their measurements were forced into ranking sets. Others immediately congregated to the non-parametric bandwagon as they hoped or perceived that these "clean" tests would act as an antidote for a poor experimental design, a poor measurement instrument, a malfunctioning research

---

[1] H. Hotelling and M. R. Pabst, "Rank Correlations and Tests of Significance Involving No Assumptions of Normality," The Annals of Mathematical Statistics, 7 (1936), pp. 39-43.

setting, or any combination of these or other disasters.

Needless to say, there was chaos in the field. Some held to the conviction that these were inferior statistical tests, to be used as infrequently as possible. Others used the nonparametric mode of analysis in research situations that could have been more appropriately analyzed using the normal curve model.

Soon, a balance seemed to emerge with the advent of computing devices. Comparative mathematical methods of comparison such as the Pitman Efficiency or Asymptotic Relative Efficiency were loaded in favor of normal theory. It was mathematically impossible to check the validity of the use of a parametric technique when the postulates upon which the particular theory was derived were blatantly violated. It was comparable to saying that the area of the rectangle was approximately length times width when the assumption of a planar surface was known to no longer be true. In comparing the relative usefulness of a parametric to a distribution-free test, it was necessary to state the exact conditions and extent of assumption violations if one wished to make a meaningful comparison of two comparable tests. It was the advent of a computer simulation under a variety of exacting conditions that was to enable statisticians to make a judgmental decision in the comparison of statistical tests.

Geary stated that "Normality is a myth; there never was, and never will be, a normal distribution."[1] It was not until

---

[1] R. E. Geary, "Testing for Normality," Biometrika, 34 (1947), pp. 209-42.

the use of computer investigations into the effect of violations of parametric assumptions that one could hope to arrive at a proper prospective about the appropriateness of the normal model. As is usually the case, truth lies at neither extreme. Both classes of statistical testing have their proper niche' in the research field. The question is under what specific set of conditions is the practitioner most correct in making a choice between comparable modes of statistical analysis.

Specific Points of Comparison

It is incorrect to state that either a distribution-free or a parametric test is superior to its counterpart. Each of these classes of tests have definite points of superiority over the other. Bradley outlines several specific comparisons that could be considered in contrasting two tests.[1]

In terms of simplicity of derivation, the nonparametric tests require a mathematical sophistication that is far less than that required for the normal curve statistics. Parametric tests, if they are to be understood at a level beyond that of "cookbook" comprehension, demand a mathematical background of integral calculus and advanced probability theory. Since many researchers who use these tests are not prepared to such a degree, their understanding could be viewed as an exercise in blind faith, and their choice of a statistical test is based upon whatever is currently fashionable and acceptable in their field of study.

---

[1] Bradley, pp. 17-24.

On the other hand, most distribution-free tests require a mathematical background no more powerful than that of many high school graduates. With the knowledge of simple probability and combinational formulae, one can readily understand the meaning of his computed results. One of the serious weaknesses of research in different fields is the incorrect or inexact use of statistics in the data analysis. At least with the distribution-free tests, a researcher without mathematical expertise is able to refer to the result as a simple probability ratio.

Another advantage of the distribution-free tests is their speed of application. However, with the development of sophisticated calculators and statistical computer packages, the question of speed is no longer as important as it was several years ago. Before the advent of the computer, the computation involved in using parametric statistics was an exacting and time-consuming task.

An important advantage to consider is that of statistical efficiency. Under nonparametric conditions, the distribution-free statistics may, in many cases, prove to be statistically more powerful than their corresponding parametric tests. In general, the advantage is held by the distribution-free tests whenever sample sizes are small or moderate. However, parametric tests regain power superiority in many instances when sample size increases. The concept of sample size effect will be further elaborated in the subsequent discussion of statistical power.

When both types of tests are analyzed under the stringent conditions required for the parametric tests, the relative advantages of the normal curve tests is very slight under many population distributions.

The parametric tests have a definite advantage when it comes to the choice of significance levels. In these tests, the distribution of the test statistic, when the null hypothesis is true, is continuous. The user of a classical test has an exacting critical point for any given alpha-level. For example, if a researcher is using an alpha-level of 0.05, he is able to state exactly how much of a deviation from the hypothesized population mean will occur five percent of the time by chance factors alone. An exact alpha-level is not possible when using a statistical test that is not derived from a continuous distribution. When using a distribution-free test, the researcher is forced to use one of the finite number of alpha-levels computed as a ratio of a definite number of counts to the total number of possible combinations. If this approach is unsatisfactory, it is then necessary to use an alpha-level inexactly. The critical value of rejecting the null hypothesis at the 0.05 level actually corresponds to a critical value that is less than or equal to 0.05. Problems can occur when one hopes to equate replications of the same study using different sample sizes. Since critical values are based upon a finite number of arrangements, it is highly unlikely that any two experiments with different n's can have exacting alpha-levels. However, as sample sizes increase, this

difference between actual and theoretical alpha-levels will, for all practical purposes, be insignificant.

One of the most obvious points of contrast between the two branches of statistical testing is the types of statistics that are actually testable. Parametric tests analyze statistics that are functions of observations and measurements that have undergone some arithmetical operation, whereas distribution-free tests interpret order relationships or discrete frequency counts. As a result of this Normal Mystique, means and standard deviations have come to be regarded as ideals. Yet it must be understood that these statistics can be considered as ideal only when they are viewed as measures of symmetrical distributions at least, or normal distributions at best. The median and the range, although lacking the mathematical sophistication of its parametric counterparts, have definite utility, if only in their intrinsic clarity to those who are not quantitatively orientated.

One of the severe weaknesses of the distribution-free tests is the lack of a generally accepted methodology of testing certain higher-order complex interactions. The methods derived so far to analyze these interrelationships are generally based upon limiting cases and therefore are not truly "distribution-free". The parametric Analysis of Variance does not exhibit this weakness.

The effects of sample sizes are important in selecting between parametric and nonparametric tests. When sample sizes are extremely small, the distribution-free tests are very

efficient, highly reliable, and are only slightly less powerful than the classical tests even when contrasted under the conditions of the normal tests. In cases involving samples of thirty or more, especially for symmetrically distributed populations, the distribution of the test statistic rapidly approaches a Normal distribution due to the effect of the Central-Limit Theorem. Also, for larger samples, many nonparametric tests are computed using the Normal Curve as a limiting case. In these situations, there appears to be no distinct advantage in resorting to the distribution-free tests. It is in the intermediate sample size range, from ten to thirty, and in particular whenever the population distribution deviates greatly from normality, that the question of which class of statistical tests should be utilized is most open to debate.

Finally, a definite advantage of the distribution-free tests is their superiority in scope of application. Since they are based upon far fewer assumptions and restrictions, they are able to be applied to a far larger class of populations and in situations where population characteristics are unknown.

It is sometimes difficult to detect violations of assumptions of the classical tests. An example of one such violation is the case of a population distribution that is generally normal in shape with the exception of a concentration of cases in a far extreme location in one of the tails. Even though those ordinates in the extreme region are far greater

than that of the normal distribution, the probability of their occurrence is still quite slight in relation to the total population. Therefore, it is highly unlikely that many of these cases will be sampled in experiments of small or moderate size. Using the classical normal curve tests for such a skewed distribution would result in conclusions that could be highly suspect.

It is evident that a researcher is rarely confronted with data that perfectly conforms to the assumptions upon which a given test is based. The gap between theory and practice can never totally be bridged, with the final question coming down to one of degree. How flexible are the assumptions? At what level of violation does the test in question become less efficient than some other possible choice? These are not easily answered from a mathematical perspective. Even a slight violation of one assumption destroys the intrinsic logic of the mathematical thought process. Each individual situation must be independently analyzed, and them some truth might emerge through an investigation of each specific violation.

Robustness

Statisticians use the term robustness of a test to refer to a test's sensitivity to distortions of various kinds.[1] There appears to be no formal definition of this concept that is generally acceptable. In layman's terms, though, robustness

---

[1] Hubert M. Blalock, Jr., Social Statistics (New York: McGraw-Hill, 1972), p. 270.

means the ability to disregard certain assumptions or conditions upon which the test was theoretically derived and still maintain a sense of confidence in the result. One would have to say that a given test was robust against certain violations under specific conditions. Even the above statement gives no mathematical measure of robustness. If a study claimed that the t-test for two independent samples was robust against the violation of the normal curve assumption for sample sizes greater than thirty, one could not be guaranteed that the t-test would also be robust against the normality assumption if the additional requirement of heterogeneity of variances was violated. In trying to generalize about robustness, there is the serious difficulty that these conditions and factors tend to be highly interactive.[1] If two separate experiments showed that a test was highly robust under two different conditions, there would be no guarantee that this same test would again be robust under both conditions acting simultaneously. Bradley states that

> . . . it is clear that in order to convert 'The ---- test is robust against the ---- assumption' into a meaningful and accurate statement about robustness, it would have to be accompanied by a quantitative definition of robustness, a complete quantitative statement of the degree or extent of the violation, and a complete specification of the exact sampling and test conditions under which the test is to be performed. If all this were done, the statement would be so particularistic as to have little general appeal, which perhaps explains why the type of statement quoted survives in its amorphous,

---

[1] Bradley, p. 27.

> undefined, unqualified form. It also explains why that form is so completely inaccurate and so utterly meaningless.[1]

Statistical Power

The dictionary defines power as the ability or capacity to perform effectively. Statisticians use the same concept in an analogous fashion. Technically, the power of a statistical test is defined as the probability of rejecting the null hypothesis when it is, in fact, false. Symbolically, power can be specified as follows:

Power = 1 - (probability of a Type-II error).

Statistical power is a measure of performance. It is a quantitative statement as to whether or not a given test is able to detect a significant difference that is actually present. Power is a measure of a test's sensitivity, a measure of that test's potential to correctly reject a statement of no difference. If a major role of a statistical test is to help detect important trends, then it should logically follow that statistical power can be used as a gauge to measure whether or not a given test is adequately performing its function. Since power is expressed as a probability statement, it can be conceptualized as a percent. For example, a statement that the power of a given test is 0.569 means that the test in question will correctly reject the null hypothesis 56.9 percent of the time and incorrectly accept the statement of no difference 43.1 percent of the time. Since the power of a test is dependent upon a false null

---

[1] Bradley, p. 43.

hypothesis, this power cannot be calculated with prior knowledge of the population parameter under the research hypothesis, H_1.

Runyon and Haber have stated that statistical power is a function of five different factors:

1) sample size
2) alpha level
3) the nature of H_1, the alternate hypothesis
4) the use of correlated measures, and
5) the nature of the statistical test.[1]

Statistical power is a direct function of sample size, since an increase in sample size will increase the power of a given test if all other factors remain constant.

Power is also related to the alpha-level, the probability of rejecting a true null hypothesis. As the alpha-level is lowered, the likelihood of a Type-I error is decreased. As a result, there is a subsequent increase in the probability of a Type-II error. Since statistical power and the probability of a Type-II error are complementary events, any increase in a Type-II error probability will result in a decrease in power. It therefore follows that a lowering of the alpha-level for a given experiment will also lower the power of a statistical test.

The nature of the research hypothesis also has an effect upon power. Since power increases with an increase of the alpha-level, it follows that power should increase as the critical value of the test statistic decreases. The critical

---

[1] Richard P. Runyon and Audrey Haber, Fundamentals of Behavioral Statistics (Reading, Massachusetts: Addison-Wesley Publishing Co., 1976).

value of the test statistic is closer to the mean of the sampling distribution in a one-tailed, directional test than it is in a two-tailed, non-directional test. Therefore, a directional statistical test will be more powerful than a non-directional test.

Correlated measures have a definite effect upon power. When two sets of subjects are matched on a variable that highly correlates with the criterion variable, the use of a statistical test that takes correlation into account will be more powerful than a test which does not take advantage of error reduction.

Finally, the nature of the test itself must be considered. As a general rule, parametric tests are more powerful than their nonparametric counterparts when the assumption underlying the use of the parametric tests are valid. When certain assumptions are violated, as mentioned earlier, it is not necessarily the case. The reason behind this difference in power is the fact that the parametric tests make optimal use of all information that is available when the populations from which the samples are derived are normally distributed. When a set of scores is analyzed using the normal assumption, the magnitudes of their differences are taken into account. In many distribution-free tests, these scores are converted to simple ranking arrangement. Thus, order only, and not their absolute counts, is measured by the nonparametric methods. The loss of statistical information by ranking renders the distribution-free tests less powerful when samples are drawn from normal populations.

Since statistical power appears to be an efficient method of measuring the effectiveness of a given test's ability to detect differences between groups, it would also appear to be a satisfactory point from which to proceed to form comparisons between two or more tests operating under identical circumstances and assumptions.

Efficiency is a term used to compare the power of two comparable tests. It is concerned with the increase in sample size that is required to make one test as powerful as the other test to which the first one is being compared. It is generally expressed as a percent using the following formula:[1]

$$\text{Power Efficiency of Test B} = 100(N_a/N_b), \text{ where } N_b$$

represents the sample size required to make Test B as powerful as Test A when Test A's power is computed from sample size N_a.

The most common measure of power efficiency used to compare the relative power of a distribution-free test to that of its parametric counterpart is the Pitman Efficiency, also known as the Asymptotic Relative Efficiency.[2] This Asymptotic Relative Efficiency, commonly abbreviated as the A.R.E., is designed to give a quantitative value of relative power as the sample sizes approach infinity. Due to the effects of the Central-Limit Theorem, the A.R.E. ratio would always have a

---

[1] Runyon and Haber, p. 319.

[2] Gottfrield E. Noethar, "On a Theorem of Pitman," The Annals of Mathematical Statistics, 26 (1955), pp. 64-8.

limiting value of one as sample sizes became larger. In order to avoid the uninterpretable approach to unity, the A.R.E. uses the techniques of differential calculus to interpret the final result. Simply, the Asymptotic Relative Efficiency is the limiting case of the ratios of the second derivatives of the two power functions under consideration.

This Pitman Efficiency conveys a precise mathematical interpretation for power comparisons, but is not without certain questionable aspects. Primarily, the A.R.E. is calculated under conditions which meet all of the assumptions of the parametric test, while giving no comparable advantage to the distribution-free test that is being compared.[1] From a practical perspective, no one is interested in relative power at a point called "infinity". It is not to say that the Asymptotic Relative Efficiency is a useless comparative measure for parametric and distribution-free tests. The A.R.E. gives some perspective to begin a comparison. It should not be used, however, as an ultimate piece of evidence in discussing the relative worth of two different tests. The Asymptotic Relative Efficiency does, however, tend to be a lower bound for relative efficiency in the neighborhood of H_o.

Power efficiency is an inappropriate measure to compare the power of two tests at a consistent sample size. As a result, the Relative Power Efficiency can be used in these situations. For a given coordinate on a power function graph

---

[1] Bradley, p. 59.

(measure of Shift), the Relative Power Difference of Test A compared to Test B is defined as $P_a - P_b$, where P_a and P_b are the respective power estimates for the two tests.

The distribution-free tests tend to be most powerful for extremely small sample sizes ($n < 10$). As sample sizes become larger ($n > 30$), the effects of the Central-Limit Theorem give the power advantages to the classical tests under most population distributions. Employing moderate sample sizes between this range of ten to thirty, the question of relative power advantage is generally unanswered.

The Monte Carlo Method

If one wished to investigate a power comparison between a nonparametric and a parametric test under a sampling condition of moderate size and under a population that represented a rather drastic violation of the normal curve, a purely mathematical approach would be impossible. In discussing this problem, Scheffe' states that ". . . we realize that standards of rigor possible in deducing a mathematical theory generally cannot be maintained in deriving the consequences of departure from these assumptions."[1] It is therefore necessary to turn to a simulation experiment, or more correctly, to a Monte Carlo simulation.

The Monte Carlo Method is defined in the broad sense as

---

[1] Henry Scheffe', The Analysis of Variance (New York: John Wiley, 1959), p. 331.

any technique for the solution of a model using random numbers or pseudorandom numbers.[1] Random numbers are those numbers that exhibit the properties of uniformity and independence. Uniformity means that in the long run each number will appear an equal number of times. Independence requires that the next digit is not predictable from inspection of the previous ones. That is to say, there is no pattern in the repetition of the digits. A table of random digits could be generated by placing ten numbered balls into a jar and then drawing and recording each individual draw, making sure that the balls were adequately mixed after each draw and replacement. Most tables of random numbers are those generated by applying a deterministic algebraic formula which results in numbers that for all practical purposes are considered to behave as random numbers.[2] In developing a set of pseudorandom numbers, one must determine that the properties of independence and uniformity are not violated.

The Monte Carlo Method has its origins through the work of the American mathematician, Stanislaw Ulam. While trying to determine the probability of completing a game of solitaire to the final card, Ulam decided to attempt to find a solution using simulations.

> When I could not devise a general solution, it occurred to me that the problem could be examined heuristically, that is, in a way

---

[1] Jack P.C. Kleijnen, Statistical Techniques in Simulation, Part I (New York: Marcel Dekker, Inc., 1974), p. 6.

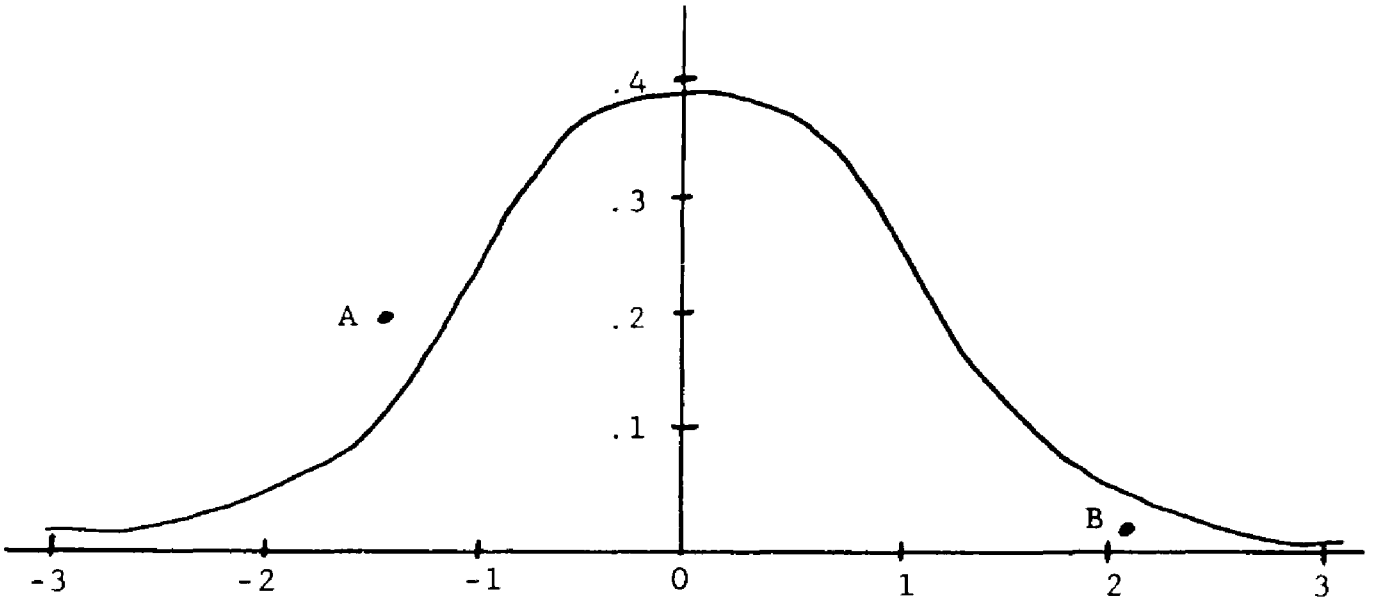
[2] Kleijnen, p. 8.

> that the examination would at least give an idea of the solution. This would involve actually playing out a number of games, say 100 or 200, and simply recording the results. It was an ideal task for a computer and was at the origin of the Monte Carlo Method.[1]

The Monte Carlo technique can be used to simulate the drawing of random samples from a specified population shape. Suppose a random sample of size ten was required from the unit normal curve, $f(x) = Ke^{(-x^2/2)}$, where $K \doteq 0.399$, and $f(x)$ represents the frequency at point x. A five-digit random number could be selected to represent a possible sample from this population. The first three digits would represent the x-value on the curve (some theoretical score) ranging from -5 to +5. The first digit, from 0 to 9, would represent the integral region along the x axis. To simulate the drawing of negative values, random digits larger than 500 could represent x-values that are negative according to some predetermined rule. The digit 561 could stand for -0.61, 603 for -1.03, 724 for -2.24, etc. The fourth and fifth digits would represent the f(x) value. Assume that the five-digit number that happened to be generated was 64519. This would correspond to an ordered pair on the graph of (-1.45, 0.19). (See Figure 1)

---

[1] Stanislaw Ulam, Mathematical Thinking in the Behavioral Sciences (New York: W. C. Freeman, 1968), p. 166.



$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{(-x^2/2)}$$

FIGURE 1

AN EXAMPLE OF THE MONTE CARLO METHOD IN ESTIMATING AN AREA UNDER A UNIT NORMAL CURVE

Call this point A. Upon plotting A on the graph, one would notice that point A does not fall underneath the curve f(x). In Monte Carlo terminology, this is classified as a "miss". This point is disregarded and another five-digit random number is selected. If the random number happened to be 21402, then point B would be plotted as the ordered pair (2.14, 0.02). The second ordered pair is under the curve and is classified as a "hit". A "hit" represents an individual selected for the sample. The first person sampled had a score of 2.14. The procedure is continued in a similar fashion

until the required sample size is obtained. Utilizing either pre-programmed or user-written computer programs, very complex sampling arrangements can be simulated using Monte Carlo techniques. With the development of these methods, the answers to certain questions can be obtained in a simulation manner that cannot be tested directly by mathematical methods.

"Student" t-test

The most commonly used parametric test for determining the probability that two samples are drawn from identical populations is the t-test for two independent samples. This particular test is frequently used for research designs that are simple Experimental-Control group studies. The various forms of the test were developed by the statistician, W. S. Gossett, who wrote under the pseudonym of "Student".[1] Gossett was employed by the Guiness Brewery in Dublin, Ireland. Since it was corporate policy that no employee of the brewery be allowed to publish any research findings under their own name, Gossett adopted the pen name of "Student". There are three major assumptions required for the proper use of t-test.[2] The level of measurement of the dependent variable must be at least from an interval scale. The samples must be randomly selected and drawn from populations that are normally distributed

---

[1] Morris Hamburg, Statistical Analysis for Decision Making (New York: Harcourt, Brace and World, Inc., 1970), p. 344.

[2] Blalock, p. 223.

and of equal variance. The t-distribution has the following form:

$$f(t) = c(1 + \frac{t^2}{v})^{-(v + 1)/2}$$

where

$$t = \frac{\bar{x} - u_x}{s_{\bar{x}}},$$

c = a constant required to make the area under the curve equal to unity, and

v = the degrees of freedom.

The variable t ranges from negative infinity to positive infinity. The constant c is actually a function of v, so that for a particular value of v, the distribution of f(t) is completely specified. Thus, f(t) is a family of functions, one for each value of v. As with the standard normal distribution, the t-distribution is symmetrical and has a mean of zero. The variance of the t-distribution is greater than one, but the variance approaches one as the sample sizes and the degrees of freedom become larger. Thus, the variance of the t-distribution approaches the variance of the standard normal distribution as the sample sizes increase. It can be shown mathematically that the unit normal curve is the limiting case for the t-distribution as the degrees of freedom approach infinity. This approach is very rapid, and many researchers will use the standard normal curve when the degrees of freedom are greater than thirty, although technically, the t-distribution is the correct functional form.

It should be understood that the t-distribution makes

use of a special case of the Central-Limit Theorem. In its general form, the Central-Limit Theorem relates the distribution of a sample mean to the population mean. However, the theorem can also be applied to the difference between any two independent sample means. It states simply that the distribution of the differences between two sample means will also approach the normal distribution as the size of these samples becomes larger. The *t*-distribution actually provides a correction factor for the normal curve *z*-test. The correction is necessary whenever sample sizes are small and the population standard deviation is unknown. The *t*-test flattens out the normal curve, pushing the critical region out further from the hypothesized mean for the distribution. It provides a necessary margin of error to keep the probability of a Type-I error at acceptable levels. As mentioned earlier, the "Student" *t*-test was originally developed with the mathematical assumption that the observed samples were drawn from a normally distributed population. When the actual population parameters are unknown, serious loss of power can occur when the population frequency distribution greatly deviates from absolute normality. In certain cases of moderate sample size, the "Student" *t*-test for two independent samples does not appear to be the ideal choice for a two sample test for Shift, the traditional form of a simple Experimental-Control group study.

The Mann-Whitney U-Test

The most popular distribution-free alternative to the "Student" t-test is Wilcoxon's rank sum test. Equivalent forms of his testing procedure appeared in the literature under various names, the most widely recognized version being known as the Mann-Whitney U-Test. An initial form of the test was first introduced by Wilcoxon[1], who used S, the sum of the ranks assigned to population 1, as the test statistic. Wilcoxon's test was extended to the case of unequal sample sizes by White[2] and van der Reyden. It was, however, the work of Mann and Whitney[3] that popularized this test.

There are four major assumptions underlying the use of the Mann-Whitney U-Test:

1) Both samples are randomly selected from their respective populations,
2) In addition to independence within each sample, there is mutual independence between the two samples,
3) Both samples consist of continuous random variables, and

---

[1] F. Wilcoxon, "Individual Comparisons by Ranking Methods," Biometrics, 1 (1945), pp. 83-3.

[2] C. White, "The Use of Ranks in a Test of Significance For Comparing Two Treatments," Biometrics, 8 (1952), pp. 33-41.

[3] H. B. Mann and D. R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other," The Annals of Mathematical Statistics, 18 (1947), pp. 50-60.

4) The measurement scale is at least ordinal.[1]

If we let E and C represent two populations, then the null hypothesis for the Mann-Whitney U-Test is that A and B have the same distribution. In a directional sense, the alternative hypothesis, H_1, would state that E is larger than C. We would accept the alternative hypothesis if the probability of a score from E being larger than a score from C is greater than one-half. If e represents one observation from E and c represents one observation from C, H_1 would be written as H_1: $p(e > c) > \frac{1}{2}$.

The test statistic U for the Mann-Whitney version of the test is given by the number of times that a score from group E precedes each score from group C. Suppose E contains the observations 9, 11, 15, 20 and C contains 6, 8, 9, 7, 13. Arrange these scores in an increasing order as follows:

$$6_c,\ 7_c,\ 8_c,\ 9_e,\ 10_c,\ 11_e,\ 13_c,\ 15_e m\ 20_e,$$

Observe the score 10. Only one E score, namely 9_e, precedes 10_c. A value of one is recorded. Two E scores, 9_e and 11_e, precede 13_c. An additional value of two is recorded. Therefore, the test statistic $U = 1 + 2 = 3$. An alternative statistic, U', can be measured by proceeding in the opposite direction. U' becomes 17. U' can also be computed from U by the following relation:

$U' = U - (n_c n_e)$, where n_c represents the number of cases in the sample taken from C, and n_e being the number of

---

[1] W. J. Conover, Practical Nonparametric Statistics (New York: John Wiley, 1971), p. 224.

cases in the sample taken from E. In the event of tied scores, correction factors have been determined that have the result of making the test slightly more conservative. However, in its theoretical development, the Mann-Whitney U-Test assumes that the scores represent a distribution which has underlying continuity.[1]

It has been shown that as sample sizes become larger than twenty, the sampling distribution of U rapidly approaches the normal distribution.[2]

$$\text{Mean} = \frac{n_e n_c}{2} \quad \text{and}$$

$$\text{Standard deviation} = ((n_e)(n_c)(n_e+n_c+1)/12)^{\frac{1}{2}}$$

The significance of an observed value of U can be determined by a z-test from the normal curve of mean zero and unit variance.

In terms of power efficiency, the Mann-Whitney U-Test is almost as powerful as the t-test under the assumption of normality. The Asymptotic Relative Efficiency of the Mann-Whitney Test as compared to the t-test is 3/pi or 0.955.[3] For a uniform distribution, the Asymptotic Relative Efficiency is 1.00, and for certain distributions, the A.R.E. can be greater than one when compared to the t-test. An interesting

---

[1] Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences (New York: McGraw-Hill, 1956), p. 123.

[2] Mann and Whitney, pp. 50-60.

[3] Conover, p. 235.

feature of the U-Test is the fact that the Asymptotic Relative Efficiency can never be less than 0.864, a valuable property shown by Hodges and Lehmann.[1]

The Mann-Whitney U-Test possesses many characteristics that make it an ideal test for Experimental-Control group situations. Conover presents the following discussion:

> Ranks may be considered preferable to the actual data for several reasons. First, if the numbers assigned to the observations have no meaning by themselves, but rather attain meaning only in an ordinal comparison with the other observations, then the numbers contain no more information than the ranks contain. Such is the nature of ordinal data. Second, even if the numbers have meaning but the distribution function is not a normal distribution function, the probability theory is usually beyond our reach when the test statistic is based on the actual data. The probability theory of statistics based on ranks is relatively simple and does not depend on the distribution in many cases. A third reason for preferring ranks is that the A.R.E. of the Mann-Whitney test is never too bad when compared with the two-sample t-test, the usual parametric counterpart. And yet the contrary is not true; the A.R.E. of the t-test compared to the Mann-Whitney test may be as small as zero, or "infinitely bad." So the Mann-Whitney is a safer test to use.[2]

Deviations From Population Normality

Parametric statistical tests have, as one of their major assumptions, the condition of normality for the parent population.

---

[1] J. L. Hodges and E. L. Lehmann, "The Efficiency of Some Nonparametric Competitors of the t-test," The Annals of Mathematical Statistics, 27 (1957), pp. 324-35.

[2] Conover, pp. 223-4.

Careful investigation of actual data, however, can lead to the obvious conclusion that the Gaussian normal curve is by no means an adequate form to describe reality. The normal relationship may well be appropriate for certain populations, but approximate normality might better be considered as the exception rather than the rule.

An early population model that was a vast deviation from normality was Allport's J-curve of conforming behavior.[1] According to his theory, social conformity is a continuously distributed variable whose greatest density appears concentrated in the region of complete conformity with an elongated tail spread over increasing degrees of nonconformity. Situations such as time of arrival of workers at a factory, length of parking in timed parking zones, and degree of conformity in church rituals are examples of behavior that appeared to follow this model. The above is one example of observed population characteristics whose frequency distribution is non-normal and whose analysis under classical parametric methods would result in severe loss of power.

It is not uncommon in many behavioral science situations for an influential, discretely distributed variable to be left uncontrolled, either because of ignorance of its existence or influence, or because of an inability to control it. If one considers the continuously distributed dependent variable, X, a greater understanding of this population distribution can be

---

[1] F. H. Allport, "The J-Curve Hypothesis of Conforming Behavior," Journal of Social Psychology, 5 (1934), pp. 143-83.

obtained by analyzing the discretely distributed causal variable, U, which is decomposed into five levels.[1] For each level of U, the dependent variable will have a separate distribution, so that the total distribution of X is a composite distribution obtained by summing ordinates of the k component distributions of U. The exact shape of the dependent variable's distribution will be determined by several factors: the k levels of U, and the area, mean, shape, and variance of each individual k level. As an example of such a distribution, Bradley cites the case of an experiment measuring the time required to reach up from a fixed position and operate a push button. The total time required to complete the assigned task was a function of the number of errors before the goal was accomplished. Each error forced the participant to start over and attempt to perform the task again. The final distribution of X was a bizarre shaped distribution, composed of k quasi-normal distributions where k was the number of errors before successful completion. The more difficult the task, the greater the number of k levels, and as k increased, there was a greater deviation in X from the normal curve model.

Careful partitioning of parent populations can show that a population in question is the composite of several sub-populations, each of which exhibit approximate normality. Researchers continually use the normal curve as the model for population distributions and compute statistical tests that

---

[1] James V. Bradley, "A Common Situation Conductive to Bizarre Distribution Shapes," The American Statistician, 31 (1977), pp. 147-50.

were developed from the normality model. It could be said that a researcher, upon discovering that the population under study exhibits such irregular forms, should simply find and control the variable responsible for it. In certain situations, that may well be the appropriate course of action. However, in behavioral science research, this method of control may be impossible or may not be of interest. If the study at hand is an Experimental-Control design, the experimenter may "not be interested in eliminating an assignable cause, but rather in coping with (i.e., drawing inferences about) a population in which it is free to vary naturally."[1] Vast areas of research contain such uncontrolled variables, and many research situations measure variables whose population shape is unknown.

Mixed-Normal Distributions

Blair and Higgins have investigated statistical power of such distributions that have appeared in social science research. They have labeled such population frequency distributions as mixed-normal distributions.[2] These are the simplest form of the above mentioned case where the causal variable, U, is decomposed into only two levels. The distribution of intelligence test scores, generally normal in shape, possesses a small lump at the lower portion of the tail. These extreme scores possibly represent those individuals whose mental

---

[1] Bradley, The American Statistician, p. 149

[2] R. Clifford Blair and James J. Higgins, "A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's t Statistic Under Various Nonnormal Distributions," Journal of Educational Statistics, 5 (1980), p. 115.

deficiencies are due to a single pathological factor, such as a genetic defect, and not the result of a chance combination of multiple factors.[1] Intelligence scores are more accurately measured by a mixed-normal distribution than by the Gaussian normal curve.

The mixed-normal distribution is the composition of two normal distributions. Depending on the composition of these two sub-populations, the final graphical form of the composite function may closely approximate the traditional normal curve or may appear in a form that is a very signficant variation from normality.

---

[1] Bradley, American Statistician, p. 149.

In their analysis, Blair and Higgins[1] used the mixed-normal distribution whose functional form was as follows:

$$f(x) = \frac{0.95}{(2\pi)^{\frac{1}{2}}} \cdot e^{-x^2/2} + \frac{0.05}{10(2\pi)^{\frac{1}{2}}} \cdot e^{-(x-33)^2/200}$$

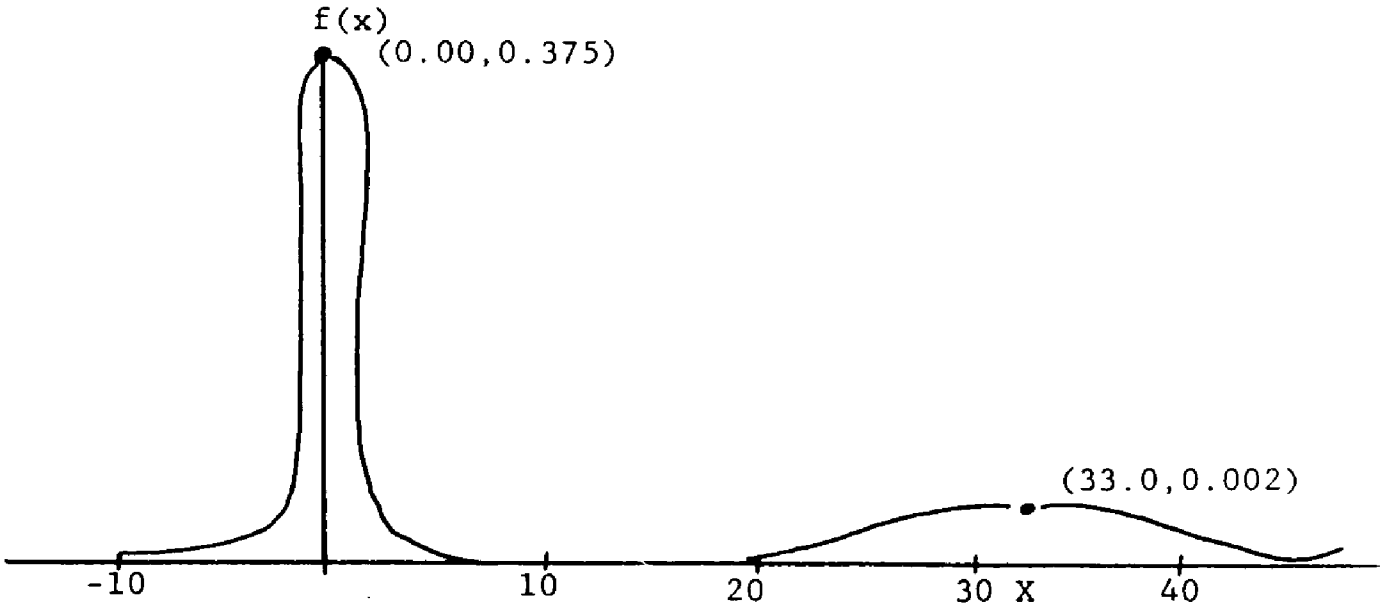


FIGURE 2

A MIXED-NORMAL DISTRIBUTION WHERE DENSITY = 0.95, SEPARATION = 33.0, AND SUB-POPULATION STANDARD DEVIATION RATIO IS 1 TO 10

The function f(x) is actually composed of two normal curves, $f_1(x)$ and $f_2(x)$, the first of which has a zero mean and unit variance while the second partition has a mean value of thirty-three and a standard deviation of ten. The constants, 0.95 and 0.05, represent weights for the proportion of cases

---

[1] Blair and Higgins, p. 315.

that fall within each of the two subdivisions of the population. The curve has interesting properties, not only because it approximates empirical data in research situations, but also because Blair and Higgins, in a Monte Carlo study of the t-test and the Mann-Whitney U-Test under mixed-normality, reported results that appear to contradict previously accepted standards of statistical power. Namely, under this population distribution, the power advantage will fluctuate between the two tests at small sample sizes and will increase in favor of the Mann-Whitney U-Test as samples increase in moderate sizes.

A generalized functional form of the mixed-normal distribution could be expressed as:

$$f(x) = \frac{DN}{\sigma_1} \cdot e^{-(x-(u + \frac{SP}{2}))^2/2\sigma_1^2} + \frac{(1-DN)}{2} \cdot e^{-(x-(u - \frac{SP}{2}))^2/2\sigma_2^2}$$

where DN - the proportion of cases in sub-population 1,

$0 \leq DN \leq 1$

1-DN = the proportion of cases in sub-population 2

u = the mean for the normal curve distribution if SP = 0

σ_1 = the standard deviation for sub-population 1

σ_2 = the standard deviation for sub-population 2

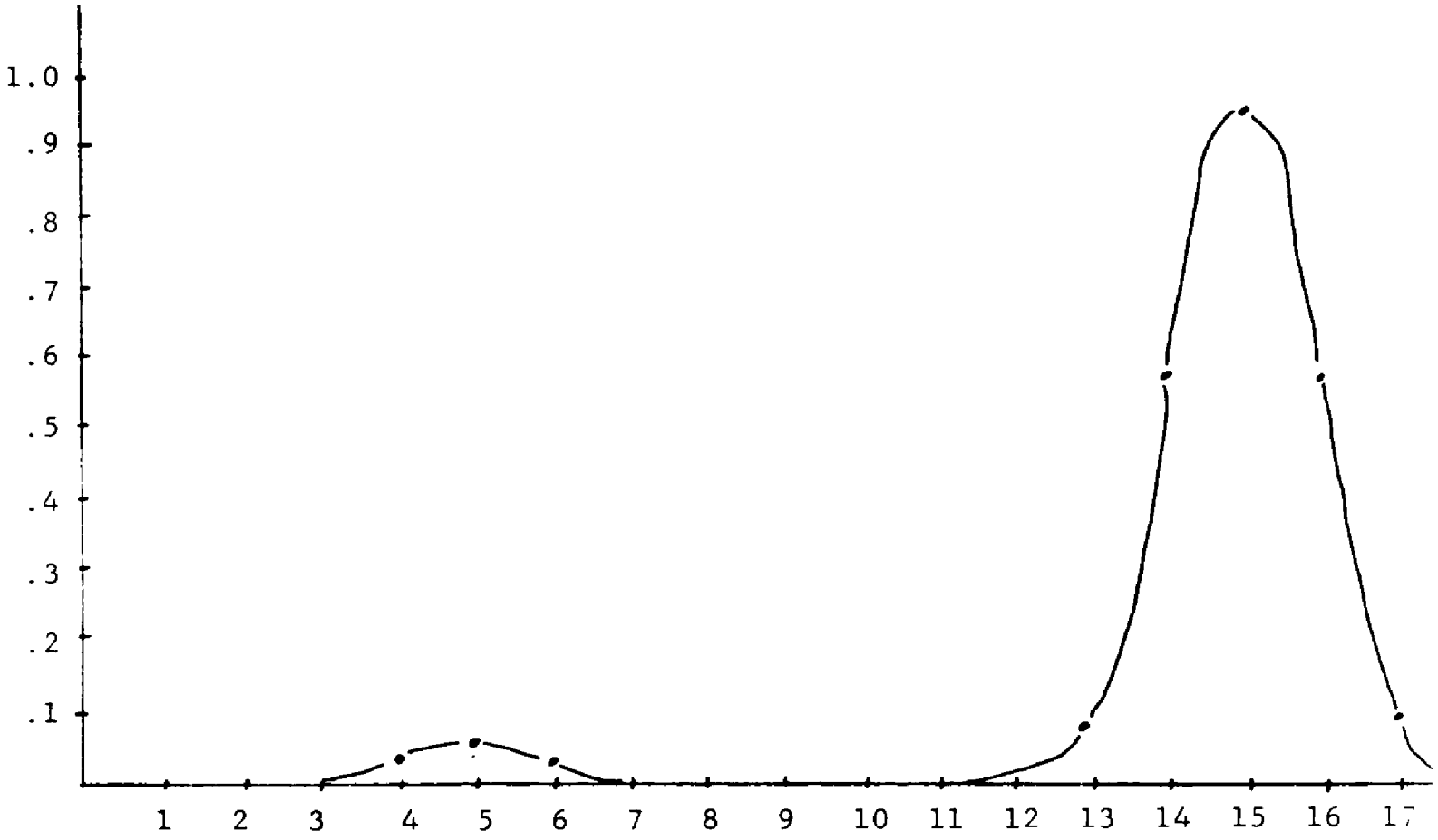
SP = a mean separation factor

x = a score on any measuring scale or instrument.

If $\sigma_1 = \sigma_2$, then the family of curves is a function of two variables, DN and SP. DN will be called the Density Factor and will be used to represent the relative proportion of cases in each of the two sub-populations. SP is the

Separation Factor, the difference between the means of the two sub-populations as measured on the variable X. The special cases in which SP=0 or in which DN=0 or 1, are simply functional representations of the unit normal curve, $f(x) = C \cdot e^{-(x^2)/2}$. Figure 3 gives an example of the mixed-normal distribution for DN=0.05 and SP=10. Appendix A gives the graphs for the twenty-five mixed-normal distributions under investigation. Appendix B tests the means, standard deviations, and coordinates for all the mixed-normal distributions under study.

The mixed-normal population model could be applicable in certain educational and behavioral science situations. A teacher is confronted many times with a class that is really two classes in one. One group may easily be able to handle the material, and another group may exhibit performance that is far removed from the first. Emotions such as hate and love, pride and shame, in fact any quantity or quality which represents a dichotomy, could possibly be fit by some form of a mixed-normal model, especially if the Separation Factor between the two sub-populations is very large.



DN = 0.05
SP = 10

U = 10
1 = 2 = 1

$$f(x) = (0.05) \cdot e^{-(x-5)^2/2} + (0.95) \cdot e^{-(x-15)^2/2}$$

FIGURE 3

A MIXED-NORMAL DISTRIBUTION IN WHICH
DENSITY IS 0.05 AND SEPARATION IS 10.0

Specific Research Questions

The purpose of this research is to conduct a Monte Carlo computer simulation on various forms of the mixed-normal distribution and to determine the effects of DN and SP on the Relative Power Difference of the Mann-Whitney U-Test as compared to the "Student" t-test for two independent samples. The Blair and Higgins' study will also be replicated with one major variation. In their study, the DN value was equal to 0.95 and SP was equal to 33. In addition, the standard deviation of the second sub-population was ten times as large as that of the first sub-population. The present study will investigate the effects of the difference in sub-population standard deviations upon the Relative Power Difference of the two tests in question.

In order to compare the results of this study with those of Blair and Higgins as well as those of Blair, Higgins, and Smitley, moderate sample size will be operationally defined as $n_1=n_2=18$.

The research questions specifically analyzed by the present study are:

1. Are the directional Type-I error rates generated by computer simulation consistent with the theoretical value of 0.05 for the t and U-tests under mixed-normal distributions of different DN and SP values?
2. In the case of moderate sample sizes, does the Mann-Whitney U-Test tend to be more powerful than

the "Student" t-test under a mixed-normal distribution for various values of DN and SP?

3. Is the Relative Power Difference of the Mann-Whitney U-Test as compared to the "Student" t-test functionally related to DN, SP, and Shift?
4. At specific values of DN, is the Relative Power Difference of the Mann-Whitney U-Test as compared to the "Student" t-test functionally related to SP and Shift?
5. At specific values of SP, is the Relative Power Difference of the Mann-Whitney U-Test as compared to the "Student" t-test functionally related to DN and Shift?
6. For the specific case of a mixed-normal distribution with DN=0.95 and SP=33, does sub-population standard deviation have any effect on the Relative Power Difference of the Mann-Whitney U-Test as compared to the "Student" t-test?

Definition of Important Terms Relevant to This Particular Study

1. Central-Limit Theorem: Let $X_1, X_2, \ldots X_n$ be a sequence of identically distributed random variables each with mean, u, and variance σ^2.

   Let $T_n = X_1 + X_2 + \ldots + X_n$

   Then for each fixed value of z, as n tends to infinity, Prob $\left(\frac{T_n - nu}{\sigma \sqrt{n}} > z\right)$ approaches the probability that the standard random variable Z exceeds z.

2. Density (DN): The proportion of cases that fall in one of two sub-populations in a mixed-normal distribution.
3. Mann-Whitney U-Test (also known as the Wilcoxon Test): A common distribution-free statistical test that estimates the probability that two samples could have been drawn from the same population. The computation of the test statistic is derived by assigning ranks to the elements of the two samples.
4. Mixed-Normal Distribution: A theoretical population distribution that is forced by the composition of two separate normal distributions.
5. Monte-Carlo Technique: A method of determining the validity of a model using random digits or pseudorandom digits.
6. Parameter: Some measure of a total population that may be estimated by a measure computed from a sample.
7. Parametric Test: A statistical test that makes a hypothesis about the value of a parameter in a statistical density function.
8. One-Tailed Test: A statistical test in which the research hypothesis predicts the direction of the difference.
9. Population: The complete set of elements having some specific measures or characteristics.
10. Power of a Statistical Test: The probability that the application of a given test results in the decision of correctly rejecting a null hypothesis that is in fact false.

11. Probability of Type-I Error: The probability of falsely rejecting a null hypothesis that is true.
12. Probability of Type-II Error: The probability of failing to reject a null hypothesis that is false.
13. Relative Power Difference of Test A in Relation to Test B: At any coordinate on a power graph,

    $P_a - P_b$, when P_a = power for Test A and
    P_b = power for Test B
14. Sample: A subset of a total population.
15. Separation (SP): The distance in standard units between the means of the two sub-populations in a mixed-normal distribution.
16. Shift (DI): The difference in standard units between the population mean of the Experimental Group and the population mean of the Control Group.

CHAPTER II

REVIEW OF RELATED RESEARCH

Introduction

As stated earlier, there is no definitive consensus as to the appropriateness of utilizing nonparametrical tests over their normal-curve counterparts. However, when one encounters research advocating or discouraging the use of these distribution-free tests, the reasoning behind such recommendations usually falls into one of two categories. First, certain studies claim that the t or the F-tests are insensitive to violations of their underlying assumptions, and as a consequence, can be used even under situations in which these assumptions are blatantly violated. Another category of research centers around the concept of statistical power, generally positing the position that either the normal-theory tests or their alternatives possess a greater probability of correctly rejecting false null hypotheses.

The Questions of Robustness

The first category of research began to emerge as early as the beginning of the 1930's. The general conclusion of earlier investigations may be summarized under the generalization that the t or F-tests are nearly "immune" to violations

of their assumptions or can easily be made "immune" if certain precautions are taken.

Norton conducted a study in which samples of F's were obtained from distributions having identical means but violating to some extent the assumptions of normality and homogeneity of variance.[1] According to theory, if the null hypothesis is true and all assumptions were met, F-values that would exceed the critical value should occur exactly 5 percent or 1 percent of the time, depending upon the alpha-level chosen. A summary of Norton's conclusion is as follows:

1. When both samples are drawn from the same population distribution, the actual shape of this distribution had very little effect upon the percentage of F ratios exceeding the theoretical limits.
2. When sampling occurred from populations having the same shape but different variances, or having different shape but the same variance, there again was little noticeable effect. Generally, the percentage of cases exceeding the theoretical 5 percent limit was between 6.5 percent and 7.0 percent, a slight inflation of Type-I error rates.
3. Problems could occur whenever different variances are combined with different shaped population

---

[1] D. W. Norton, "An Empirical Investigation of Some Effects of Non-Normality and Heterogeneity of the F-Distribution," Diss. State University of Iowa, 1952.

distributions. In some instances, there was a serious discrepancy between the theoretical and actual number of cases exceeding the critical values.

These results led Lindquist to conclude that "unless the heterogeneity of either form or variance is so extreme as to be readily apparent upon inspection of the data, the effect upon the F-distribution will probably be negligible."[1] The appropriateness of Lindquist's commentary upon Norton's work would seem to be in question unless it were coupled with an attempt to define the term "extreme".

The most frequently quoted research in defense of the parametrical invulnerability is that of C. Alan Boneau. His full series of investigations concerned the t-test for two independent samples and was based upon a Monte Carlo investigation that involved 1,000 different simulations for each variation being considered.[2] Random samples were drawn from either normal, rectangular, or exponential population shapes. Means were equal to zero and variances of either 1 or 4 were studied. For several combinations, t-tests were computed using sample sizes of 5 and 15.

As with Norton, this original study of Boneau analyzed Type-I error rates, and his findings were quite similar to

---

[1] E. F. Lindquist, Design and Analysis of Experiments in Psychology and Education (Boston, Massachusetts: Houghton Mifflin, 1953), p. 86.

[2] C. Alan Boneau, "The Effects of Violations of Assumptions Underlying the t-test," Psychological Bulletin, 57 (1960), pp. 49-64.

that of Norton. There were, however, a few particular situations that resulted in high error rates.

In sampling from two normal distributions of different variances, when sample sizes were different (n_1=5, n_2=15), 16 percent of the *t*'s fell outside the 5 percent limits. When population distributions were changed to exponential forms with equal variances, the *t*-test became conservative. At alpha-levels equal to .05, only 3.1 percent of the cases exceeded the theoretical limits for sample sizes of 5. At sample sizes of n_1=n_2=15, the corresponding percentage of *t*'s that fell into the critical region increased to 4.0 percent. For rectangular distributions of equal variances, the theoretical *t*-distribution provided an excellent fit to the empirical results. For n_1=n_2=5, 5.1 percent of the cases were beyond the critical value of 5 percent. As sample sizes increased to 15, exactly 5.0 percent of the cases exceeded the limits.

Sampling from non-normal distributions with different variances (in the ratio of 1 to 4) tended to increase Type-I error rates from sampling simulations when variances were equal, although Boneau did not simulate all possible combinations.

Sampling from different distributions for each sample tended to result in slightly higher Type-I error rates, but as sample sizes were increased, the theoretical *t*-distribution provided a closer fit.

Boneau concluded that his results show that "the approach

to normality is rather rapid since sample sizes of 15 are generally sufficient to undo most of the damage inflicted by violations of assumptions."[1] He proceeded to temper his statement with a warning that highly-skewed distributions would require "slightly larger" sample sizes. Boneau went on to say that "it would appear that the t-test is functionally a distribution-free test, providing the sample sizes are sufficiently large and equal."[2]

Bradley severely criticized Boneau and similar authors of robustness studies. Bradley stated that robustness is a highly questionable concept since there is no generally accepted quantitative definition for the concept: how much of a deviation from theoretical probability levels is acceptable before one is able to say that a test is no longer robust? He admonished Boneau and others not only for a lack of a quantitative measure for robustness but also for "insufficient qualification, the highly peculiaristic nature of the qualifying conditions and for biased presentations."[3]

Bradley proposed the following definition for robustness:[4]

$$\left| p - \text{alpha} \right| \leq \frac{\text{alpha}}{2}$$

That is to say, the test in question would be "robust" if the departure of p, the true probability of Type-I error,

---

[1] Boneau, p. 60.

[2] Boneau, p. 60.

[3] Bradley, British Journal, p. 145.

[4] Bradley, British Journal, p. 145.

from alpha, the theoretical probability level, was negligible, i.e., less than or equal to one-half of alpha. Bradley's proposal was very liberal, allowing for a sizeable deviation from the theoretical level. This means that for alpha = 0.05, a test would be classified as robust if the true probability of a Type-I error was between 0.025 and 0.075. Outside the range, a test would be non-robust.

Using this definition of robustness, Bradley conducted simulation studies for a two sample independent t-test under a population distribution that was basically L-shaped in form. He concluded that his criterion for robustness was not always met until the smallest sample size was 256, hardly evidence for the robustness of the t-test when normality is violated.[1]

Bradley also made note of the restricted range of Boneau's work. In these earlier studies, whenever non-normality was present, sample sizes were always equal. Whenever any heterogeneity of variance was present, Boneau's populations had the same shape.

Bradley's major thesis was that any statement about a test's robustness was only valid under the very specific conditions under which that particular analysis was performed. A major point was that robustness cannot be generalized as a result of a few studies under extremely stringent conditions. He also noted that the conditions of robustness are highly interactive, that particular combinations of alpha-levels, sample sizes, and population shapes may produce results

---

[1] Bradley, British Journal, p. 147.

that are non-robust for tests even if these tests would appear to be robust under conditions tested separately.

The Question of Power

The second rationale for advocating the use of parametric tests in place of the distribution-free forms is the claim that the normal curve models exhibit advantages in terms of their statistical power, their ability to reject null hypotheses that are actually false.

An early study of power efficiency under the normal curve assumption was that of Dixon.[1] His investigation assessed the power of four nonparametric tests (rank-sum, maximum deviation, median, and total number of runs) for the difference of two sample means drawn from normal populations of equal variances. Sample sizes for Dixon's study were three, four, and five. Dixon concluded that all four nonparametric tests had high power efficiency for small samples and small alpha-levels as compared to the "Student" *t*-test. As the difference between means is increased, power efficiency tended to decrease slightly. As the level of significance increased, the power superiority of the Wilcoxon test (Mann-Whitney *U*-Test) will increase slightly. At sample sizes of n=5, the Wilcoxon Test had a greater power advantage than either the median or maximum deviation tests. Its highest power advantage was for small alternatives between sample means, although at no point

---

[1] W. J. Dixon, "Power Under Normality of Several Nonparametric Tests," The Annals of Mathematical Statistics, 25 (1959), pp. 610-14.

was the power efficiency of the Wilcoxon Test greater than 0.964 or less than 0.88. The slight power superiority for the "Student" t-test was to be expected since Dixon's study was conducted under conditions of sampling from a normal distribution.

Hodges and Lehmann[1] performed Asymptotic Relative Efficiency studies on the Mann-Whitney U-Test relative to the t-test. They concluded that the Mann-Whitney U-Test (Wilcoxon Test) can have a Pitman efficiency as high as infinity but never lowr than 0.864. They concluded that "the use of the Wilcoxon Test instead of the 'Student's' t-test can never entail a serious loss of efficiency for testing against shift."[2]

Boneau conducted a Monte Carlo simulation comparing the relative power of the Mann-Whitney U-Test to the t-test.[3] In this study, he sampled again from normal, rectangular, and exponential distributions. His study generated 1,000 pairs of random samples for each comparison. Sample sizes again were 5 or 15.

Under normal distributions with honogeneous variances, Boneau discovered that the two tests in question demonstrated nearly equal power with a slight advantage to the t-test except for very low alpha-levels or large differences between means, where the t-test showed a definite trend of superiority.

---

[1] Hodges and Lehmann, pp. 324-35.

[2] Hodges and Lehmann, p. 356.

[3] C. Alan Boneau, "A Comparison of the Power of the U and t-Tests," Psychological Review, 69 (1962), pp. 246-56.

Upon situations of heterogeneous variances in the ratio of 1 to 4, there was very little effect on power for either of the two tests when samples were of equal sizes. Under the compound conditions of heterogeneous sample sizes and heterogeneous variances, the t-test maintained its power superiority.

It would appear that under all apparent conditions of normality of population distributions, the "Student" t-test held a power superiority over the Mann-Whitney test, at least as far as one could generalize from the works of Boneau.

When both samples were taken from the same non-normal distributions, either rectangular or exponential, the t-test again held its slight power advantage, except for specific cases for small differences in means, where the U-Test possessed greater power.

The situation differed, however, when one sample was drawn from an exponential distribution and the other from a normal distribution. In these analyses, there were mixed results. For simulations with sample sizes of 5, the Mann-Whitney U-Test was more powerful for most measures of shift, but with an increase of sample sizes to 15, the t-test regained the advantage for the most part.

Neave and Granger enacted simulation studies from populations that were the result of the superposition of two normal curves, a mixed-normal population distribution.[1] They employed samples of $n_1=n_2=20$ as well as $n_1=20$ and $n_2=40$.

---

[1] Henry R. Neave and C.W.J. Granger, "A Monte Carlo Study Comparing Various Two Sample Tests for Differences in Mean," Technometrics, 10 (1968), pp. 509-22.

This experiment utilized 500 pairs of samples for each point of investigation. The particular mixed-normal population curve was equivalent to one with a Density factor of DN=0.25 and a Separation factor of SP=2.00. Eight different tests were analyzed: the t-test, the U-Test, and six other nonparametric tests. At an alpha-level of 0.05, the Mann-Whitney test was superior, with the exception of cases in which the sub-population standard deviations were unequal (in the ratio of 1 and 2). Under these situations, the t-test had a very slight power advantage, 0.026 at the maximum level of differences. With alpha decreased to 0.01, the t-test was inferior in all cases except one, and even in that case by a power difference of 0.006.

Based on these larger sample sizes, Neave and Granger's work would appear to counter Boneau's conjecture that increase in sample size should result in power advantage for the t-test.

Toothaker simulated samples of sizes less than or equal to 5 from normal, uniform, and skewed distributions to compare the relative power of the U, and the t-tests.[1] This study generally supported the conclusions of Boneau's earlier work. There was little difference in power efficiency between the two tests for these small sample sizes.

Blair, Higgins, and Smitley compared the relative power of these same two tests under an exponential population

---

[1] L. E. Toothaker, A Empirical Investigation of the Permutation t-test as Compared to the Student's t-test and the Mann-Whitney U-Test (Madison, Wisconsin: Wisconsin Research and Development Center for Cognitive Learning, 1972).

distribution.[1] The authors chose this particular distribution since the Pitman Efficiency of the Mann-Whitney test as compared to its normal curve alternative was 3.0, but Boneau's research claimed little difference in power when sampling from a population whose distribution was exponential in form. Their study used a wider range of sample sizes than Boneau and increased the number of samples drawn from each analysis from 1,000 to 5,000. Sample sizes ranged from 3 to 81.

For a one-tailed test for alpha=0.05, the Mann-Whitney U-Test showed a clear power superiority. In no case did the t-test result in a higher percentage of correct rejections of null hypotheses. The percentage of rejections that favored the U statistic was as high as 33 percent for unequal n's and 31 percent for equal sample sizes. Even for small sample sizes, there was a substantial difference in power in favor of the Mann-Whitney test. The percentage increased to 37 percent for unequal n's for a two-tailed test at the 0.05 significance level. In their two-tailed analysis, the source of Boneau's apparent error becomes evident. For $n_1=n_2=6$, the Mann-Whitney test had a power advantage for local Shift differences (small differences in means), but the t-test regained superiority as differences in sample means became larger. Boneau mistook this result under very small sample size conditions in drawing his conclusions. The study by Blair, Higgins,

---

[1] R. Clifford Blair, J. J. Higgins, and William D. S. Smitley, "On the Relative Power of the U and t-tests," British Journal of Mathematical and Statistical Psychology, pp. 114-20.

and Smitley demonstrated clear power superiority for the non-parametric test under the specific population of mixed-normality distribution as sample sizes became larger.

The same pattern continued for alpha equal to 0.01. In these simulations, the U-Test achieved a power advantage as high as 43 percent for unbalanced sample sizes and as high as 32 percent for situations in which sample sizes were equal. The advantage of the "Student" t-test, again at very small sample sizes and large Shift magnitudes, never became larger than 11 percent.

Blair, Higgins, and Smitley contended that the Mann-Whitney U-Test should not be considered only to be a small sample procedure since it appears to control the Type-I error rate better than the t-test and may also exhibit power superiority. They went on to say that it was

> . . . important to realize that the optimal power properties of the t-test only apply when sampling is from a normal population and it is therefore not surprising to find that a nonparametric test proves to be more powerful than the t-test when samples are drawn from other than normal populations. The truth of this statement is clearly seen in the data presented here for the exponential distribution. After reading several statements on this subject and discussing it with several practitioners, it is the present authors' belief that a faulty line of reasoning may be the basis for this conclusion.[1]

They concluded their discussion with the statement that

> the question still remains--which statistic, Student's t or the Mann-Whitney U, is more appropriate for use with non-normal

---

[1] Blair, Higgins, and Smitley, p. 119.

> populations? The answer is, of course, distribution (and perhaps even sample size) specific. But in many research situations, little is known about the specific population shape, although there may be reasons to believe that it is not normal. Clearly Boneau's advice is of limited value in this situation. Therefore, until more is known, it seems that Hodges and Lehmann's (1956) result provides a good rule of thumb.[1]

Boneau's defense of the t-test's robustness and power superiority seems to still be accepted by many authors and researchers as a rationale for the continued use of the t-statistic even under situations where its assumptions are in question. It would now appear that more attention must be placed on specific population characteristics.

A recent extension of the above research was conducted by Blair and Higgins for one-tailed research hypotheses.[2] The authors used the same combinations of sample sizes as in the earlier research by Blair, Higgins, and Smitley. However, these simulations were conducted under six different population shapes: uniform, Laplace, half-normal, exponential, mixed-uniform, and mixed-normal distributions. Blair and Higgins were able to present four general conclusions. First, under these non-normal population shapes, the Wilcoxon statistic (the Mann-Whitney U) generally held very large power advantages over the t-statistic. Second, the theoretical measure of power comparison between two tests, the Asymptotic Relative Efficiency, tended to be a reasonably good indicator of the

---

[1] Blair, Higgins, and Smitley, p. 149.

[2] Blair and Higgins, pp. 309-35.

relative power between the two statistics under investigation. Third, results that were obtained from smaller sample simulations would tend, quite often, to be very dissimilar to results that were obtained by subsequently increasing sample sizes. Finally, the authors concluded that "because of the narrow ranges of population shapes and sample sizes investigated in some widely cited previous studies of this type, the conclusions reached in these studies must now be deemed questionable."[1]

Of interest was their analysis of a mixed-normal population distribution. The population function under their investigation was composed of two normal curves, the first of which contained 95 percent of the population Density with a unit variance. The second curve had a standard deviation of 10 and was separated from the first curve by a Separation factor of SP=33.

The authors' discussion of maximum power advantages for the two tests resulted in mixed conclusions. For small samples, there appeared to be no apparent pattern, but as sample sizes became larger, the Mann-Whitney U-Test began to demonstrate a definite superiority of power. Blair and Higgins' study was tested under four different significance levels: 0.05, 0.025, 0.01, and 0.005.

In their particular case of a mixed-normal population distribution, namely with DN=0.05 and SP=33 coupled with

---

[1] Blair and Higgins, p. 309.

sub-population standard deviations in the ratio of 1 to 10, the Mann-Whitney statistic maintained power advantage over the t-distribution with the only exceptions being certain simulations with sample sizes of 6 or less.

These results contradicted Boneau's claim that "in general the t-test is more powerful than the U-Test."[1] Blair and Higgins' work demonstrated the opposite result: namely, that the U-Test appeared to maintain a vast power advantage, especially with moderate and large sample size simulations.

Generalizations about power are difficult to defend. The question which remains unanswered is whether or not the Mann-Whitney U-Test would maintain its apparent power superiority for different forms of this mixed-normal population distribution. Relative Power Difference is a function of many factors, among which population shape seems to be a primary variable that requires further investigation.

Conclusion

The mixed-normal distribution has only been analyzed in isolated cases. Several research studies have separated conflicting results about the Type-I and Type-II error rates for the "Student" t-test and the Mann-Whitney U-Test. These studies have analyzed a wide range of population distributions. Some authors have claimed that the t-test is functionally a distribution-free test. Others have concluded from their research that the use of the t-test may result in highly distorted

---

[1] Boneau, Comparison of Power, p. 255.

results. Normally "quasi-normal" distributions are best analyzed from a parametric perspective. When the population shape deviates greatly from normality, however, the Mann-Whitney U-Test appears to hold an advantage, at least in terms of power, over the t-test. Further research is required before an answer to the following question can be found. Which of the two tests, the "Student" t-test for two independent samples or the Mann-Whitney U-Test, is a more powerful analytical tool for testing hypotheses about differences in means when the populations under investigation are of different forms of the general family of curves that have come to be known as mixed-normal distributions?

CHAPTER III

BACKGROUND AND PROCEDURES

Introduction

The general purpose of the present study is to determine which of the two inferential statistics tests in question, the parametric "Student" t-test or the nonparametric Mann-Whitney U-Test is more powerful under various forms of mixed-normal population distribution. The two tests under investigation are common in behavioral and social science research as tools in testing hypotheses concerning differences between means for Experimental-Control group studies and for other testable hypotheses. The mixed-normal population model was chosen since it represents, in certain forms, a drastic deviation from the pure normal curve distribution that is a basic theoretical assumption upon which the t-distribution is based. These mixed-normal distributions serve as models for theoretical populations. The population distrubitions are actually composed of two sub-populations, each of which are normal curves that measure some ability, performance, or trait for the two sub-groups that compose the total population.

Deficiencies in Previous Simulations and Theory

Previous studies have not analyzed these particular research questions to any great extent. Early power simulations

were limited in several aspects. Due to time and technological restraints, sample sizes were quite small, and those studies that did attempt to investigate a wider combination of sample sizes were limited in the population shapes from which the samples were drawn. No previous research attempted to explore in depth the Relative Power Difference of the t and U-test for a mixed-normal population. A few studies, however, did analyze one or two particular mixed-normal distributions without any extensive analysis of general properties for this family of population curves.

The usefulness of mathematical power models such as the Asymptotic Relative Efficiency were questionable, especially under conditions that greatly deviated from the normal model. These statistical tools, while generally providing some method of power comparisons, did not yield a quantitative result that was comparable to results found by empirical simulations. This was especially true in cases where the nonparametric test was more powerful than the t-test.

Specific Method of Simulation

This present investigation uses a Monte Carlo simulation with an Apple-II Plus microcomputer. To analyze the particular research questions under investigation, sample sizes of $n_1=n_2=18$ are chosen. The equality of Experimental and Control group size will enable comparisons to be made with recent investigations of Blair and Higgins, who chose eighteen to represent moderate sample size.

For each value of DN, the Density Factor, and SP, the

Separation Factor, random samples of thirty-six will be drawn from a particular mixed-normal population distribution. An arbitrary constant will be added to each of the eighteen experimental group scores to represent the fact that the null hypothesis of no difference between groups is false. The two tests under investigation, the parametric "Student" t-test and the nonparametric Mann-Whitney U-Test, are then computed. The number of cases for each test that results in the correct rejection of the null hypothesis is an empirical measure of power. At each difference level between means (the arbitrary constant added to each sample for the Experimental group), five hundred simulations will be performed. After five hundred trials, the value of the constant is increased by 0.25 to represent a further difference between Experimental and Control group means. The process is continued indefinitely until either the Shift factor equalled 5.75 standard units or until the distribution of the power function graphs becomes evident by inspection.

Density Factor

The Density Factor, DN, measures the proportion of the total population cases that are in the first of the two sub-populations. The particular values of DN that will be tested in this simulation are 0.50, 0.40, 0.20, 0.10, and 0.05. The wide range of DN values will represent situations in which the two sub-populations are relatively evenly divided in frequency (higher DN values) as well as situations where

one of the two sub-groups contain a far greater number of cases (lower DN values).

Separation Factors

At each particular DN value, five hundred separate simulations will be performed in which the Separation Factor, SP, is allowed to take on the values of 1, 2, 4, 6, and 10. A special case in which SP is set equal to zero will test the comparative power of the two statistical tests under normal theory, under the condition that both Experimental and Control groups are drawn from identical normal population distributions. The SP Factor measures the distance in standard units with which the means of each sub-population are removed from each other. A small SP Factor means that the two sub-population means are close to each other. In referring to real research situations, a smaller Separation Factor would represent cases in which the two sub-populations are very similar to each other in whatever trait or performance is being measured. As SP is allowed to increase, this represents a situation in which the two sub-groups that compose the total populations are very divergent in the dependent variable under study.

Analysis of Type-I Error

The special condition in which the arbitrary constant added to the experimental group is set equal to zero has additional interest. In these cases there is no difference between the two means. Here is a representation of a situation where the Null Hypothesis is true. The proportion of cases for each

test that results in the incorrect rejection of the Null Hypothesis is an empirical estimation of the Type-I error rate.

Additional Variables for Power Estimation

Any power estimation is a function of several variables in addition to the Separation and Density Factors. Two such variables are alpha-level and test direction. The hypotheses in this study will be simulated at alpha equal to 0.05 with all research hypotheses directional in nature (one-tailed).

Requirements for Valid Computer Simulations

Glass, Peckham, and Sanders presented several suggestions to be followed in computer simulation studies.[1] As far as possible, their recommendations were followed in this study.

They suggested that the computer program used to generate the random samples and to compute the statistical tests be thoroughly documented. Appendix C contains the complete program, written in Applesoft Basic, together with appropriate comments to clarify the necessary variables and procedures.

The authors also suggested that the sampled distributions should be discussed and presented as clearly as possible. Appendix B contains the population parameters and sub-population parameters for all the mixed-normal population distributions that are to be analyzed.

Any simulation study is only valid if the pseudorandom number generator is performing to optimal efficiency. Appendix

---

[1] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance," Review of Educational Research, Vol. 42 (1972), pp. 237-88.

D contains the statistical analysis of the algorithm used in Applesoft Basic to select random numbers over the range of twenty standard units required in the present study.

Glass, Peckham, and Sanders also mentioned that baseline checks of the entire simulation should be performed and reported. The special case where SP=0 is equivalent to sampling from a population distribution that is normal. The results of the normal population distribution's simulation should be consistent with previous theory. The t-test should hold a slight power advantage over the Mann-Whitney U-Test.

Method of Reporting Results

In order to answer the specific research questions in this study, the following analyses will be presented:

1) A chart of Type-I error rates as a function of DN and SP for both the U and the t-test.
2) Twenty-six charts showing the number of H_o rejections (empirical power estimates) for all twenty-five arrangements of DN and SP (as well as the normal curve situation) at various measures of Shift, i.e., difference between Experimental and Control group means.
3) Twenty-six graphs of power estimation showing power as a function of Shift.
4) A multiple regression equation will be computed with the Relative Power Difference of the Mann-Whitney U-Test as the dependent variable and DN, SP, and Shift as independent variables.

5) At each measure of Density, a separate multiple regression equation will be computed with Relative Power Difference as a function of Separation and Shift.

6) At each measure of Separation, a separate multiple regression equation will be computed with Relative Power Difference as a function of Density and Shift.

7) For the specific mixed-normal curve analyzed by Blair and Higgins,[1] a chart will be presented showing the Relative Power Difference of the Mann-Whitney U-Test as a function of the ratio of sub-population standard deviations in addition to a multiple regression equation predicting Relative Power Difference.

---

[1] Blair and Higgins, pp. 309-35.

CHAPTER IV

ANALYSIS OF DATA

Introduction

Statistical decision making is surrounded by error. In making specific decisions about hypotheses, a researcher must be concerned with the possibility of an erroneous decision from two perspectives. One of these error types could be viewed as a "radical" error; its counterpart in the opposite sense could be classified as "conservative" error. These probabilities of an incorrect decision must be balanced against each other for a true perspective of the meaning of any decision rendered from statistical reasoning.

A "radical" error is one in which a researcher claims to have discovered a significant finding, when in fact, the observed difference is due to chance factors. Take for example the experiment of tossing a simple coin and using the recorded results to answer the question as to whether or not the coin might have been two-headed. One might suggest that the logical solution to the problem would be to simply examine the coin and state one's conclusion with absolute certainty. Analyzing the coin itself may be a fine approach in the coin-tossing problem. It is analogous, however, to measuring the total population in a statistical experiment. If access were possible to total populations, then the exercise of

statistical inference through measuring samples from the population under investigation would be a series of meaningless trials from a practical perspective. Total populations are beyond the resources of most experimenters. One must be satisfied with the hazardous methods of making decisions based on the results of conclusions drawn from only a subset of the total group.

If it were agreed that the decision on the coin problem would be based on the results of ten tosses, it is conceivable, though highly unlikely, that ten consecutive heads would appear, even though the coin is one-headed.

In statistical thinking, one must acknowledge that about one time in a thousand trials of tossing ten one-headed coins the result will appear as ten consecutive heads and no tails. Using the observed information and claiming that the coin is biased is an error in the "radical" sense. It is a claim that a meaningful difference truly exists (that the coin is different from the ordinary coin) when the opposite would be discovered if the researcher had the opportunity to examine the coin instead of passing judgment on the basis of a limited number of trials.

Such an incorrect decision is referred to as a Type-I error, the probability of rejecting the Null Hypothesis of no significant difference when the reality of the situation is that the statement of no difference should be accepted. It is a claim of having observed a meaningful deviation that does not really exist. In the Experimental Control group testing

situation, the measured difference may simply reflect the laws of chance from sampling identical populations.

A researcher sets limits beyond which he is willing to take a chance and state that differences exist, knowing only too well that there is always some probability that the difference is due to a rare sampling of divergent elements from two populations that are identical.

Analysis of Type-I Error Rates

The present research sets the probability of Type-I error at 0.05, meaning that in the long run an incorrect decision of rejecting the Null Hypothesis will occur five times in every one hundred experiments.

The first research question attempts to investigate the accuracy of this 5 percent error rate when samples are drawn from different mixed-normal populations. The decisions about rejecting the Null Hypothesis are made by the use of the "Student" t-test for two independent samples and the Mann-Whitney U-Test.

Table I presents the results of the five hundred simulations at each arrangement of Density (DN) and Separation (SP) for the family of mixed-normal curves where the Shift Factor (DI) was zero. Setting DI = 0 represents the situation where there is no difference between the means of the Experimental and Control group.

The overall Type-I error rate for the five hundred total simulations is 0.049 for the t-test and 0.050 for the Mann-Whitney U-Test. These results are remarkably consistent

TABLE I

SIMULATED TYPE-I ERROR RATES

(Theoretical alpha level = 0.05)

Density (DN)	Separation (SP) 0.0		1.0		2.0		4.0		6.0		10.0	
	t	U	t	U	t	U	t	U	t	U	t	U
0.05	.046	.046	.064	.052	.044	.052	.046	.072	.050	.046	.024	.036
0.10			.054	.052	.042	.038	.076	.080	.070	.050	.048	.042
0.20			.052	.054	.058	.058	.066	.052	.032	.024	.042	.048
0.40			.042	.048	.052	.046	.072	.062	.036	.040	.038	.044
0.50			.064	.062	.038	.044	.042	.038	.040	.054	.044	.064

Mean Type-I Error Rates: t-test = .049

U-test = .050

with the theoretical value of 0.05 for both tests. Under the condition of sampling from a normal curve model, the probability of Type-I error is 0.046 for both tests. For the twenty-five mixed-normal simulations, the t-test has a higher error rate in twelve cases, the U-Test has a higher error rate in another twelve cases, and in one specific case, when DN = 20.0 and SP = 2.0, both tests have identical rates of 0.058. There appears to be no discernible pattern as to the effect of Density or Separation on the tendency of one test or the other to be more accurate than its counterpart in terms of Type-I error.

Using Bradley's definition of robustness, namely that a test is robust if the actual probability of Type-I error is within the range of $(1/2)\cdot(\text{alpha})$ in either direction of the theoretical alpha-level, only three of the twenty-five mixed-normal cases fail to attain this robust status. In the particular simulation where Density = 0.05 and Separation = 10.0, the "Student" t-test has a Type-I error rate of 0.024. The computed result fails to meet Bradley's standard by only 0.001, based on the range of acceptability of 0.025 to 0.075 for the theoretical alpha-level of 0.05. At Density = 0.20 and Separation = 6.0, the Mann-Whitney U-Test also fails to attain a robust classification, again by the margin of 0.001. At only one location, when Density = 0.10 and Separation = 4.0, do both tests fail to meet the criterion. The t-test exceeds its allowable error rate by 0.001 and the U-Test by 0.005.

It is clear that from the perspective of Type-I error

there is no advantage in choosing one test over the other. The probability of falsely rejecting a true Null Hypothesis is, for all practical purposes, identical for both tests under investigation. Identical in the context does not mean consistent throughout all values of Density and Separation, but rather identical in the sense that without the knowledge of exact population frequency distribution, the best estimate that can be made is that the error rate for Type-I errors is similar for both the "Student" t-test and the Mann-Whitney U-Test.

The Estimation of Power

The question of "conservative" error, the Type-II error, is the major direction of this study. Type-II error is the probability of accepting a Null Hypothesis that is, in fact, false. It is the probability of stating, in the Experimental-Control group context, that no significant differences exist between the two groups under study when these samples are actually drawn from populations whose means were not identical.

Statistical power is the complement of the probability of a Type-II error. Power represents the probability of making the right decision to reject a false Null Hypothesis, to correctly state that there is a significant difference between the two groups.

This study investigates whether the "Student" t-test or the Mann-Whitney U-test is a more powerful test for detecting differences between means when the population under study is mixed-normal in shape. Since the term mixed-normal

represents a limitless number of different curves, twenty-five specific population forms were analyzed for different values of Density (DN) and Separation (SP). Appendix E includes the results of twenty-six simulations--the twenty-five different mixed-normal distributions as well as a power study for the special case of the generalized mixed-normal model that represents a normal population curve. Here is found the power estimates for both the U and the t-test as well as their difference of power, known as the Relative Power Difference of the Mann-Whitney U-Test as compared to the "Student" t-test (PWRDIFF). These values of PWRDIFF are recorded for increments of 0.25 for Shift (DI), the difference between means for the Experimental and Control group. Appendix F contains the twenty-six power function graphs for both tests. These graphs give a visual perspective of the comparable power of each test for varying values of Density, Separation, and Shift.

Table II shows which of the two tests possess power superiority for the different values of DN and SP. Analysis of Table II indicates that for small measures of Separation, the parametric t-test maintains a definite power advantage regardless of the Density Factor. When the Separation Factor equals 1.0 or 2.0, the t-test is the more powerful test. However, when the Separation Factor becomes 4.0 or larger, a different pattern emerges. The Mann-Whitney U-Test is generally more powerful at all measures of Density whenever the Separation Factor exceeds the value of 4.0.

An interesting arrangement occurs in simulations where

both the Density and Separation Factors become large. The larger Density Factors, those approaching 0.50, represent situations in which the sub-populations are relatively equal to size. The larger Separation values depict situations in which the two sub-populations that compose the total population are far removed from each other. At these levels, neither test has consistent power superiority. The notation U→t indicates a change in power advantage as the Shift (DI) Factor, the difference between actual population means, is increased. Careful analyses of these particular simulations show that the t-test becomes the more powerful test only at relatively large values of Shift.

Table III presents an analysis of those specific mixed-normal populations that do not maintain a consistent advantage for one test over the other. The Point of Intersection in Table III represents that value of Shift at which the U-Test initially loses power advantage. The Maximum Power Difference shows the strength of the t-test's superiority as well as the location of Shift at which the maximum advantage occurs. With the exception of simulations in which Separation = 4.0, this Point of Intersection occurs at extremely large measures of Shift. A large Shift Factor represents populations in which there is a large (three or more standard units) difference between Experimental and Control group means. As a result, the estimated power for both tests is very high at these points. The power advantage for the t-test is never larger than 0.128, and these points of

TABLE II

POWER SUPERIORITY AT DIFFERENT COMBINATIONS OF DN AND SP FOR MIXED-NORMAL POPULATION DISTRIBUTIONS

		DN				
		0.05	0.10	0.20	0.40	0.50
SP	1.0	t	t	t	t	t
	2.0	t	t	t	t	t
	4.0	U	U	U	U→t	U→t
	6.0	U	U	U→t	U→t	U→t
	10.0	U	U	U→t	U→t	U→t

TABLE III

POINTS OF INTERSECTION FOR MIXED-NORMAL DISTRIBUTIONS THAT DO NOT POSSESS CONSISTENT POWER ADVANTAGES

(DN, SP)	Point of Intersection (DI)	U-Power at This Point	Maximum Power Difference For t-test	
			PWRDIFF	DI
(0.40, 4.0)	2.00	0.786	-0.056	2.25
(0.50, 4.0)	1.75	0.694	-0.072	2.75
(0.20, 6.0)	3.75	0.970	-0.010	4.50
(0.40, 6.0)	3.00	0.860	-0.084	3.75
(0.50, 6.0)	3.00	0.788	-0.128	4.00
(0.20, 10.0)	5.25	0.960	-0.016	5.75
(0.40, 10.0)	5.00	0.882	-0.070	5.50
(0.50, 10.0)	4.75	0.834	-0.124	5.75

maximum advantage occur at extremely large measures of Shift. In these particular simulations, the Mann-Whitney U-Test's superiority is extremely large at more realistic Shift values. The advantage in power is as large as 0.546 when Density = 0.20, Separation = 10.0, and Shift = 1.75.

Simulations in which there are confusing power results do not realistically detract from the U-Test's advantage for the larger Separation values. The Points of Intersection at which the t-test regains the power advantage appear at the further extremes of the power function graphs.

Greatest confusion occurs when Separation equals 4.00. At some value less than 4.00 for Separation, the t-test loses its power advantage. At all simulations in which Separation has a value less than 4.00, the t-test has power superiority. At Separation values greater than 4.00, the U-Test is, for the most part, a superior test in terms of power. It would appear that at some point of Separation between two and four standard units there is a value of Separation at which the U-Test begins to dominate and continues to dominate. The U-Test advantage is definite for small measures of Density, but it becomes uncertain at Density values that represent near equality in frequency between the two sub-populations in both the Experimental and Control groups.

Table IV shows the location of maximum power superiority for both tests at the different values of Shift. Table IV clearly depicts a high power superiority for the Mann-Whitney U-Test under specific population conditions.

TABLE IV

MAXIMUM POWER ADVANTAGES FOR THE
t AND U TESTS AT DIFFERENT MEASURES OF SHIFT

Shift	t	(DN, SP)	U	(DN, SP)
0.25	.028	(.05, 1)	.060	(.20, 6)
0.50	.044	(.50, 1)	.164	(.10, 10)
0.75	.040	(.40, 1) (.40, 2)	.310	(.10, 10)
1.00	.062	(.50, 2)	.394	(.10, 10)
1.25	.044	(.50, 2)	.484	(.10, 10)
1.50	.044	(.50, 2)	.532	(.10, 10)
1.75	.026	(.50, 4)	.546	(.20, 10)
2.00	.040	(.40, 4)	.514	(.20, 10)
2.25	.056	(.40, 4)	.480	(.20, 10)
2.50	.060	(.50, 4)	.432	(.20, 10)
2.75	.072	(.50, 4)	.408	(.50, 10)
3.00	.040	(.40, 4)	.344	(.50, 10)
3.25	.062	(.40, 6)	.300	(.40, 10)
3.50	.080	(.50, 6)	.266	(.50, 10)
3.75	.110	(.50, 6)	.232	(.40, 10)
4.00	.072	(.40, 6)	.158	(.40, 10)
4.25	.133	(.50, 6)	.086	(.50, 10)
4.50	.094	(.50, 6)	.068	(.20, 10)
4.75	.060	(.50, 6)	.040	(.20, 10)
5.00	.048	(.50, 6)	.028	(.20, 10)

At most moderate measures of Shift, the advantage for the U-Test is great, being as high as 0.546. On the other hand, the advantage for the t-test never becomes larger than 0.080 except in the previously mentioned situations in which the Shift Factor takes on a very large value.

The second research question is a complex one. There is a definite difference in power between the two tests under the mixed-normal population model. Knowledge of Separation and Density is essential if one wishes to state whether the "Student" t-test or the Mann-Whitney U-Test is the more appropriate choice in testing for differences between means of mixed-normal populations.

Functional Relationships Between Variables

The third research question under investigation is the functional relationship between the major component variables. The Relative Power Difference (PWRDIFF) of the Mann-Whitney U-Test as compared to the "Student" t-test is the dependent variable. This variable PWRDIFF is the difference in power estimation between the two tests at given values for the independent variables, Density (DN), Separation (SP), and Shift (DI). The starting point for analysis will be the zero-order as well as the partial correlation coefficients for the variables in question. Table V lists the zero-order correlation coefficients, and Table VI presents the first and second-order partial correlation coefficients.

By far, the highest correlation exists between the Relative Power Difference and Separation. The zero-order

TABLE V

ZERO-ORDER CORRELATION COEFFICIENTS
FOR THE ENTIRE STUDY

	PWRDIFF	DI	DN	SP
PWRDIFF	1.000	-0.106*	-0.105*	0.609**
DI		1.000	0.258**	0.451**
DN			1.000	0.089
SP				1.000

* = Significant at .05 Level

** = Significant at .01 Level

TABLE VI

PARTIAL CORRELATION COEFFICIENTS
FOR THE ENTIRE STUDY

Independent Variable	Dependent Variable	Variables Controlled	Correlation
PWRDIFF	DN	SP	-0.202**
PWRDIFF	DN	DI	-0.081
PWRDIFF	DN	SP, DI	-0.086
PWRDIFF	SP	DN	0.625**
PWRDIFF	SP	DI	0.740**
PWRDIFF	SP	DN, DI	0.741**
PWRDIFF	DI	SP	-0.538**
PWRDIFF	DI	DN	-0.082
PWRDIFF	DI	SP, DN	-0.514**

** = Significant at .01 Level

correlation between these two variables is 0.609. This is evident by a simple examination of the simulation results. The positive direction of the correlation demonstrates that the power advantage rapidly increases in favor of the U-Test as the difference between the two sub-population means (the Separation Factor) increases. An analysis of the second-order partial correlation coefficient between PWRDIFF and SP with the effects of DN and DI being removed confirms the strength of the Separation Factor, since the partial correlation coefficient increases to r = 0.741. Clearly, the Separation Factor between sub-population means has a significant effect upon the difference in power between the two tests. A greater difference between the sub-population means results in a greater tendency for the U-Test to be the more powerful test. The exception to such a conclusion at very large (and very unrealistic) values of Shift has already been noted.

The zero-order correlation between Relative Power Difference and Density is -0.105. Relative Power Difference takes on a negative value whenever the t-test is more powerful. When the effect of Separation, highly correlated in a positive sense with PWRDIFF, is removed, the correlation increases in the negative direction to a value of -0.202.

When the effect of Shift is removed from the correlation between Relative Power Difference and Density, the negative relationship approaches zero, indicating that the influence of Density on the t-test's power superiority is only evident at larger Shift values. When the effect of Density is statistically

adjusted, it becomes apparent that there is no significant relationship between PWRDIFF and DN.

The complex interaction of the variables is again emphasized by observing the zero-order correlation between Relative Power Difference and Shift. The zero-order value is -0.106. There is a very slight decrease in absolute strength when Density's effects are removed. However, when Separation's influence is statistically removed, the value of the first-order partial correlation coefficients drastically increases in a negative sense to -0.538. The second-order partial correlation coefficient, controlling for both Separation and Density, is -0.514. In situations that require a very large difference (Shift) between Experimental and Control group means before the simulation is completed, the t-test has superiority. The t-test holds the power advantage only for larger values of Density when Separation is also large. When the effect of this Density Factor is removed, the partial correlation between PWRDIFF and DI reduces to a nonsignificant value of -0.082.

Of these three factors, DN, SP, and DI, the effect of SP is most significant. The strong positive correlation between PWRDIFF and SP shows the U-Test's superiority for large differences between sub-population means. Density's influence becomes a relatively minor factor when other effects are statistically removed. The Shift Factor (DI) correlates with the t-test's superiority, indicated by negative values of PWRDIFF. When Density's effect is controlled, this correlation

also approaches zero.

There also exists a significant positive correlation between Shift and both Density and Separation. An analysis of these correlations does not present any meaningful insight into the relationship between the variables. It is simply a result of the design of the simulation. Large measures of Density and Separation represent deviations from normality. As a result, larger differences between Experimental and Control group means (Shift) are necessary before the required power values are reached that will result in the completion of a particular simulation. Values close to zero for both DN and SP will result in power estimations that quickly approach 1.00. In these cases, large measures of Shift are not required to complete the simulation. Only for simulations that deviate greatly from the normal model are the larger differences between group means required.

Finally, the correlation ($r = 0.089$) between Density and Separation is only due to the arbitrary values chosen for these variables.

As mentioned earlier, the most significant factor in predicting Relative Power Difference is that of Separation. In most situations, a large Separation Factor indicates a power advantage for the Mann-Whitney U-Test. Under these conditions, it would appear that a researcher is more likely to correctly reject a false Null Hypothesis under a mixed-normal population model if the Mann-Whitney U-Test is selected over the "Student" t-test. The U-Test is correct whenever the population

is composed of divergent elements that differ greatly in the trait, skill, or attribute being measured.

A multiple regression equation is presented for the variables under study. The Relative Power Difference (PWRDIFF) for the U-Test as compared to the t-test is the dependent variable. Density (DN), Separation (SP), and Shift (DI), as well as the interaction effects are the independent variables. Table VII presents the summary table for the multiple regression equation with hierarchical inclusion variables. The column entitled "R^2 Change" gives the best perspective of the influence of each subsequent variable entered into the equation. The R^2 value gives the proportion of total variation in Relative Power Difference that can be explained by the variable in question as well as all preceding variables. The most dramatic increase in explained variation occurs at Step 2 in which the Separation Factor is included in the equation. The zero-order r between PWRDIFF and SP is 0.6094, meaning that 37.13 percent (r^2 = 0.3713) of the variation in Relative Power Difference is able to be explained by the Separation Factor acting alone.

The inclusion of Shift (DI) at Step 4 increases the R^2 factor by 0.1596. The effect of the negative correlation is emphasized by the negative regression coefficients for both first-order interaction terms involving DI. A simultaneous increase in Shift with either Density or Separation, especially at the larger values of Shift, may frequently result in a decrease in PWRDIFF.

TABLE VII

SUMMARY TABLE FOR MULTIPLE REGRESSION ANALYSIS FOR THE ENTIRE STUDY

Order of Inclusion	Independent Variable	Multiple R	R^2	R^2 Change	Simple r	Standardized Regression Coefficient	Unstandardized Regression Coefficient
1	(DN)	.1049	.0110	.0110	-.1049	-.0438	-.0535
2	(SP)	.6300	.3969	.3859	.6094	.0365	.8525
3	(DN)(SP)	.6303	.3973	.0004	.2852	.0123	.0126
4	(DI)	.7463	.5569	.1596	-.1060	-.0086	-.0825
5	(DN)·(DI)	.7480	.5594	.0025	-.1636	-.0572	-.2580
6	(SP)·(DI)	.7515	.5647	.0053	.1424	-.0035	-.3456
7	(DN)·(SP)·(DI)	.7515	.5648	.0001	.0224	.0022	.0863
	Regression Constant	-----	-----	-----	------	------	-.0435

The inclusion of all these independent variables as well as interaction terms gives a multiple R value of 0.7515. The R^2 value of 0.5648 means that 56.48 percent of the variation in Relative Power Difference can be explained by the variables in this study. The value $1-R^2$, or .4352, indicates that 43.52 percent of the variation in the dependent variable is due to factors other than those considered here.

The Analysis of Density

A deeper understanding of the effects of the components on power estimation will require separate analyses at each level of the two major independent variables, Density and Separation. The first such variable to be composed is that of Density (DN), the proportion of the total population that lies in each sub-population. Tables VIII, IX, and X present the results for the decomposition on levels of Density. At every level, Separation is the major factor in predicting the Relative Power Difference between the two tests. Larger measures of Separation will result consistently in a power advantage for the U-Test whenever Density is 0.05 or 0.10.

However, a change occurs when Density takes on a value of 0.20 or larger. The zero-order correlation between PWRDIFF and DI becomes negative for Density values beyond 0.20, again indicating that for these simulations a large measure of Shift will result in a change in direction for the Relative Power Difference. At small measures of Separation, the t-test is superior. At larger values of Separation, the U-Test has a power advantage, only to lose its superiority at extreme

TABLE VIII

CORRELATION COEFFICIENTS FOR DIFFERENT LEVELS OF DENSITY

	PWRDIFF with			
	SP	DI	SP.DI	DI.SP
DN				
0.05	.6460**	.0364	.6983**	-.3493**
0.10	.7584**	.1075	.7787**	-.2901**
0.20	.6138**	-.0941	.7926**	-.6394**
0.40	.5777**	-.1616	.7408**	-.5835**
0.50	.5563**	-.1895	.7202**	-.5727**

** = Significant at .01 Level

TABLE IX

MULTIPLE REGRESSION EQUATION COEFFICIENTS IN PREDICTING RELATIVE POWER DIFFERENCE (PWRDIFF) OF MANN-WHITNEY U-TEST IN RELATION TO "STUDENT" t-test FOR THE DIFFERENCE BETWEEN TWO INDEPENDENT MEANS FOR VARIOUS VALUES OF DENSITY (DN)

Density (DN)	Separation (SP)	Shift (DI)	Interaction (SP)·(DI)	Constant	R^2
0.05	,0330	-.0065	-.0062	-.0564	.50750
0.10	.0394	-.0328	-.0027	-.0508	.61095
0.20	.0496	-.0291	-.0043	-.0532	.64289
0.40	.0409	-.0352	-.0022	-.0572	.56427
0.50	.0386	-.0344	-.0021	-.0595	.54023

TABLE X

MULTIPLE R^2 FACTOR IN PREDICTING RELATIVE POWER DIFFERENCE (PWRDIFF) FOR DIFFERENT VALUES OF DENSITY (DN)

Density (DN)	Separation (SP)	Shift (DI)	(SP)+(DI)	(SP)+(DI)+(SP·DI)
0.05	.41727	.00133	.48838	.50750
0.10	.57517	.01156	.61093	.61095
0.20	.37679	.00886	.63154	.64289
0.40	.33379	.02612	.56061	.56427
0.50	.30951	.03952	.53601	.54023

measures of Shift.

For the larger levels of Density, representing nearly equal concentrations within each sub-population, the change of power advantage occurs at smaller measures of Shift when Separation is moderately large. At large levels of Separation combined with large levels of Density, the change in power advantage takes place at extreme measures of Shift where both the t and U-test have high power estimates.

The problem area appears evident. The smaller measures of Density, 0.05 or 0.10, exhibit a consistent pattern. An initial advantage for the t-test gives way to the Mann-Whitney U-Test as the Separation factor increases. At values of Density greater than or equal to 0.20 a confusing picture emerges. The t-test's advantage for small Separation values gives way to a conflicting pattern, especially when Separation is 4.0. The analysis of different levels of Separation follows.

Analysis of Separation

Throughout the entire set of simulations, the Separation Factor between sub-population means is the most critical factor in the Relative Power Difference between the two tests. At different levels of Separation, however, the advantage is mixed between the two tests. When Separation is small, at values of 1.00 or 2.00, the mixed-normal population frequency curve is, for all practical purposes, normal. The Density Factor does not influence the shape of the population distribution. Under small Separation values, the sub-populations do not become distinguishable as separate entities. A visual

inspection of the graphs (see Appendix A) for all values of low Separation gives the impression of near-perfect normality. As would be expected in these situations, the "Student" t-test holds a power advantage, although small, in these "quasi-normal" distributions. The positive correlation between PWRDIFF and DI at these lower levels of Separation indicates that the power of the U-Test approaches that of the t-test as Shift increases. The power estimation of both tests rapidly approach 1.00 as the Shift Factor between the two groups becomes large.

At the Separation value of 2.00, there is a significant negative correlation between Relative Power Difference and Density. When the Separation Factor is set equal to 2.00, the t-test's advantage tends to become more pronounced as the two sub-population Densities approach each other (DN = 0.50).

When Separation takes on larger values, the population shapes begin to show marked deviation from normality. It is under these particular population distributions that the power question becomes less clear. When Separation is 4.00, both Density and Shift become significant predictors of Relative Power Difference. These correlations between DN and DI in relation to PWRDIFF are both negative, indicating a relative advantage to the t-test at larger values. At the Separation value of 4.00, the U-Test is more powerful at most of the cases tested. However, as Density and Shift increase, the U-Test begins to lose its initial power superiority, and at certain points of these power function graphs, the t-test regains the advantage over the U-Test. These points of intersection will

vary for different Density values when Separation is 4.00. If Density takes on the values of 0.05, 0.10, or 0.20, the Mann-Whitney U-Test is consistently the more powerful test. At larger Density values for this particular Separation Factor, there is an intersection in the power functions of the two tests. When Density is 0.40, the t-test becomes more powerful for Shift values greater than 1.75. At the Density value of 0.50, the U-Test loses power superiority after the Shift value becomes 1.50. For larger measures of Shift, the t-test's advantage becomes more pronounced until both power functions approach the value of 1.00 at the extreme Shift values.

When Separation increases beyond the value of 4.00, Density's influence begins to lessen, although it is still significant when Separation is 6.00. When Separation is exactly 4.00, the U-Test is generally a much more powerful test when Density's value is 0.20 or less. At Density equalling 0.20, an intersection of the power functions does occur at a large Shift value (DI greater than 3.50). However, the power for both tests in question is very close to 1.00, and the power advantage for the "Student" t-test is extremely slight.

When Separation is 6.00 and Density is 0.40 or 0.50, the point of power function intersection occurs at a smaller measure of Shift (DI greater than 2.75). These particular mixed-normal distributions will present difficulty in making a choice between the two tests since the interactive effect

is evident at more realistic values of Shift. The U-Test holds the advantage for small Shift values, but it loses superiority as the difference between Experimental and Control group means become larger. After the point of intersection, the Relative Power Difference is as large as -0.128 in favor of the t-test. The relative superiority for the t-test again decreases as the power of both tests approach 1.00.

The maximum Separation tested has a value of 10.0. With the extremely large difference between sub-population means, it is highly unlikely that these results would have much practical application. It is difficult to conceive of a real population whose sub-groups could differ by ten standard units. It is interesting to observe, however, the vast advantage in power for the U-Test in nearly every case. The Relative Power Advantage for the U-Test is as high as 0.546. At no point is the t-test more powerful when Density is 0.05 or 0.10. Even when Density is 0.20, the t-test possesses a very small advantage only in extreme Shift values beyond 5.00. As the two sub-populations approach each other in frequency (Density values of 0.40 and 0.50), the t-test's superiority is only apparent at Shift values greater than 4.50. Clearly, at a very large measure of Separation, the Mann-Whitney U-Test holds a distinct power advantage.

Analysis of the Blair and Higgins Mixed-Normal Curve

The final research question concerns the specific

TABLE XI

CORRELATIONS FOR DIFFERENT LEVELS OF SEPARATION

	PWRDIFF with			
	DN	DI	DN.DI	DI.DN
SP				
1.0	.0547	.4705**	.0039	.4680**
2.0	-.3900**	.1393	-.4209**	.2200
4.0	-.5508**	-.5175**	-.4495**	-.4016**
6.0	-.2998**	-.6671**	-.1600	-.6371**
10.0	-.1153	-.5298	.0215	-.5209**

** = Significant at .01 Level

TABLE XII

MULTIPLE REGRESSION EQUATION COEFFICIENTS IN PREDICTING RELATIVE POWER DIFFERENCE (PWRDIFF) FOR MANN-WHITNEY U-TEST IN RELATION TO "STUDENT" t-TEST FOR THE DIFFERENCE BETWEEN TWO INDEPENDENT MEANS AT VARIOUS VALUES OF SEPARATION (SP)

Separation (SP)	Density (DN)	Shift (DI)	Interaction (DN)·(DI)	Constant	R^2
1.0	.0142	.0154	-.0131	-.0298	.22991
2.0	-.0242	.0094	-.0136	.0140	.20035
4.0	-.1972	-.0368	-.0366	.1225	.42336
6.0	.0162	-.0320	-.0441	.1624	.46732
10.0	.1230	-.0450	-.0428	.3299	.28452

TABLE XIII

MULTIPLE R^2 FACTOR IN PREDICTING RELATIVE POWER DIFFERENCE (PWRDIFF) FOR DIFFERENT VALUES OF SEPARATION (SP)

Separation (SP)	Density (DN)	Shift (DI)	(DN)+(DI)	(DN)+(DI)+(DN·DI)
1.0	.00300	.22140	.22141	.22991
2.0	.15211	.01941	.19317	.20035
4.0	.30339	.26785	.41577	.42336
6.0	.08991	.44508	.45929	.46732
10.0	.01329	.28064	.28097	.28452

mixed-normal distribution analyzed by Blair and Higgins[1] in their recent work on statistical power. The Separation Factor was 33.0 standard units and the Density Factor was 0.95 (equivalent to DN = 0.05). Their work represents a model in which the vast majority of cases (95 percent) were located in one sub-population. The remaining 5 percent of the cases were found in an extension of the curve whose mean was separated from the mean of the first sub-population by a distance of 33.0 standard units. An additional variable was entered into the study. The sub-population standard deviations were unequal. The standard deviation of the first sub-population was 10.0 while the second sub-population had a standard deviation of 1.0. The final research question attempts to determine whether a large difference in sub-population standard deviations has any effect on their conclusion that the Mann-Whitney U-Test holds a power advantage over the "Student" t-test.

Blair and Higgins' work was reexamined by one hundred simulations at each value of Shift for the two power functions. The ratio of these sub-population standard deviations (DEVRATIO) was tested at the values of 1 to 10, 1 to 4, 1 to 2, and 1 to 1. Table XIV presents the correlations between the variables in this section, and Table XV gives the hierarchical multiple regression analysis.

The Shift Factor (DI) is the highest predictor of the Relative Power Difference (PWRDIFF) for the U-Test. As Shift

---

[1] Blair and Higgins, Journal of Educational Statistics, p. 309-35.

increases, as the difference between means of the Experimental and Control group becomes larger, the power advantage of the U-Test also increases. The Shift Factor was only limited to values of 4.00 and less. The extreme tails of the power functions, where both U and t-test power estimates would approach the value of 1.00, were not analyzed. Despite these limitations, it was apparent that the U-Test's advantage was dominant. In no instance did the t-test show greater power than the Mann-Whitney U-Test. The significant positive correlation between PWRDIFF and DI demonstrates the increase in power advantage for the U-Test as the difference between Experimental and Control group means is allowed to become larger.

The key variable under consideration, the sub-population standard deviation ratio (DEVRATIO), has a very low correlation with Relative Power Difference. The U-Test appears to have an even greater advantage at the moderate DEVRATIO values of 1 to 2 and 1 to 4. The advantage for the U-Test tends to lessen when the standard deviations are equal or when they possess extreme values in the ratio of 1 to 10. When DEVRATIO is included in multiple regression equation to predict PWRDIFF, the resulting increase in the R^2 value is only 0.003.

The answer to the last research question, at least under the limits of the present study, is quite evident. The ratio of sub-population standard deviations has little effect on the Relative Power Difference for the Mann-Whitney U-Test as compared to the "Student" t-test under the particular mixed-normal population distribution when Density is 0.95 and Separation is

33.0. The power advantage for the U-Test is very large for all different sub-population standard deviation ratios that were tested.

TABLE XIV

CORRELATIONS FOR THE MIXED-NORMAL POPULATION DISTRIBUTION WITH DN = .95 AND SP = 33.0

Variables	Correlation
PWRDIFF with	
a) DI	.8324**
b) DI Controlling for DEVRATIO	.8335**
c) DEVRATIO	.0532
d) DEVRATIO Controlling for DI	.0960

** Significant at .01 Level

TABLE XV

MULTIPLE REGRESSION ANALYSIS FOR RELATIVE POWER DIFFERENCE FOR U-TEST AS COMPARED TO t-TEST

Density = .95
Separation = 33.0

Order of Inclusion	Variable Entered	Simple r	Multiple R	R^2	R^2 Change
1	DI (Shift)	.832	.832	.693	.693
2	DEVRATIO (St. Dev. Ratio)	.053	.834	.696	.003
3	DEVDI (Interaction term)	.450	.840	.706	.010

PWRDIFF = (.115)DI + (.0963) DEVRATIO + (-.0353) DEVDI + (.1584)

CHAPTER V

SUMMARY AND CONCLUSIONS

Type-I Error

Over the last thirty years, the pendulum has been swinging, so to speak, between the advocacy of the parametric and nonparametric modes of statistical analysis. The major intent of the present study has been to stop the pendulum for a moment in time and to look at one particular question to which there is no answer: which of the two tests, the "Student" t or the Mann-Whitney U, is the "best". The word "best" in a statistical sense could mean less subject to error, less likely to make a judgemental error concerning a specific hypothesis. There are two types of inferential error, and on one of these, the two tests were both excellent. The probability of Type-I error as determined by computer simulation was nearly identical to the theoretical value for both tests. On the question of the other type of error, the results were extremely diverse between the two tests.

The major question, which of the two tests is a more powerful test, remains unanswered and will always remain so because of the peculiar nature of statistical power. The estimation of power requires an excursion into the impossible, the examination of total populations before specific samples

are drawn. Any attempt to discuss statistical power is nothing but an exercise in abstract reasoning. Yet it is a necessary exercise if one hopes to obtain any comparable standards by which to evaluate this theoretical efficiency.

Conclusions on Power

Statistical power is a function of many variables. Any realistic attempt to analyze power requires several assumptions and several arbitrary choices of variables that must be held constant or statistically removed. Five specific research questions centered around the effects of three variables, Density, Separation, and Shift on the Relative Power Difference between the U and the t-test. A fourth variable, sub-population standard deviation ratios, was given slight attention in one specific analysis. The population shape was mixed-normal, a term used to describe a population frequency distribution that was composed of two distinct normal distributions.

The Separation Factor is, by far, the most important variable in choosing the test that appears to be most powerful. When Separation is slight, when the population model is normal or so close to normal that the deviation from normality is hardly evident, the traditional choice of those who advocate the use of the parametric test is the correct one. Under normality or "quasi-normality", the "Student" t-test is more powerful than the Mann-Whitney U-Test, although the power advantage for the t-test is extremely slight. When the Separation Factor takes on values between two and four standard

units, the t-test begins to lose its advantage. For larger values of Separation, the Mann-Whitney U-Test has a very distinct power advantage, although exceptions do occur when observing large differences between Experimental and Control group means.

The Density Factor taken alone contributed little information into the question of relative power. When taken in conjunction with knowledge of Separation, however, higher Density values, representing near equality of sub-population concentration, tend to indicate a confusion about power superiority. These higher Density values, especially at the Separation level of 4.00 units, does not enable one to state that one test or the other is generally more powerful. Under these particular combinations of Density and Separation, no test has definite superiority.

Sub-population standard deviation ratio had little effect for the very specific case analyzed here. For the large measure of Separation of 33.0, the alteration of these standard deviation values accounted for very little variation in the dependent variable, the Relative Power Difference.

The most practical question that could be asked is which of the two tests should be recommended for use if one suspects that the population is mixed-normal in shape. From the results of the present study, it is obvious that the key variable in the decision is the Separation of sub-population means. If a researcher is testing a mixed-normal population whose sub-population means are more than two standard units apart,

a correct decision in rejecting a Null Hypothesis that is false is more likely to be made if the choice of a statistical test is that of the Mann-Whitney U-Test. Yet in reality, what populations would fit this model? Only one in which there is a great discrepancy between sub-groups. In educational research, possibly an example might be a situation in which the researcher was comparing results from two classrooms that were composed of sub-groups of "regular" and mainstreamed students. Even under a population distribution of obvious mixed-normality, the U-Test would only be the more correct choice in terms of power if the researcher has reason to believe that the sub-groups themselves differed in the trait being measured by a large difference of more than two standard deviation units. Another example might center around a study involving intact classrooms in which there was a large difference between sexes on the variable in question. An example of this might be some measure of physical strength.

The mixed-normal model yields results in favor of the U-Test in many cases. However, the power advantage of the t-test for Separation values less than or equal to 2.00 makes it very difficult to defend the choice of the nonparametric alternative in all but a few extreme situations. Unless one has a strong basis to reject the assumption of normality or even the so-called "quasi-normality", it would appear that the t-test is the more logical choice. The parametric alternative has a slight power advantage under population distributions of this form.

The U-Test maintains a vast power advantage in many cases of mixed-normality. From a theoretical perspective this is quite interesting. However, in making a choice between the parametric and nonparametric test, a researcher would have to hypothesize the existence of a very distinct Separation Factor before it would appear that the choice of the Mann-Whitney U-Test would be less likely to result in a Type-II error.

Contradictions in Theory

It is interesting to note the overt contradictions that occur many times between the theory on power and the actual results based on simulation studies. The Asymptotic Relative Efficiency, the most widespread measure of comparable power, presents a distorted perspective in many cases. The A.R.E. is based on several unrealistic assumptions, yet it continues as a yardstick for power comparisons. Under one particular mixed-normal population, this Asymptotic Relative Efficiency has a value as high as 45.0. It appears logically absurd to conclude that a t-test would require sample sizes forty-five times as large as the U-Test before there would be an equality of power.

Perhaps some new theoretical standard of comparable power should be developed that presents a different perspective between any two tests in question. The use of the Asymptotic Relative Efficiency can give an incorrect indication since it is based on normal theory.

Limitations of This Study

The limitations of this study are obvious. The sample size of each group was set at eighteen. The alpha-level selected was 0.05, and the Research Hypotheses were directional. As is apparent from other power simulation studies, there is no guarantee that the results would be similar if any of these factors of sample size, alpha-level, and directionality were changed. The Separation and Density values that were tested were totally arbitrary, but the emerging patterns of power give the impression that little new information would be obtained by a more exhaustive choice of Density or Separation Factors.

What emerges is a series of endless questions. Although it is true that generalizability is not possible in power simulations, the effect of Separation and the peculiar pattern of Density and Separation interaction at values of sub-population Density gives a direction for future research.

Directions for Possible Future Research

One question that would appear to be of interest is the exact or near-exact location of Separation at which the *t*-test loses its power advantage to the *U*-Test. At some point of Separation between 2.00 and 4.00, the Mann-Whitney *U*-Test becomes more powerful than the "Student" *t*-test. The location of the intersection point may, of course, be different for varying choices of sample size, alpha-level, or other variables. In the present study where the alpha-level was 0.05 and sample sizes for both groups was set at 18, the point of power change

was not able to be detected. The problem is further compounded with the interactive effects of Density. Another study similar to the present one in which Separation was allowed to range from 2.00 to 4.00 at various Density values would help to clarify the problem. Only with several extensive studies for different sample sizes, different alpha-levels, and different arrangements of Density and Separation can one hope to begin to determine at what point of deviation from normality does the U-Test become a more powerful test for differences between group means. Even with all these potential simulations, the results would only be valid for the particular population model of mixed-normality. There clearly is some point at which the Gossett model of adjusted normality, the t-distribution, no longer has a power advantage over a distribution-free analysis. The question of determining the point of deviation has never been investigated to any large extent. Several simulation studies have been performed, but most seem to center around specific population models, and their results simply report which of the two tests is more powerful. It would appear that an elaborate investigation is needed before a researcher has the information to determine the correct choice between the two statistical tests.

Conclusion

No mathematical model is perfect and no statistical test can provide infallible answers. There will always be error. A researcher can only hope to choose a model of analysis that appears to be most appropriate for the particular

problem at hand. Mark Twain once claimed that there were three kinds of lies--lies, damn lies, and statistics. The improper use and incorrect interpretation of statistics in many areas help to lessen confidence in what statistics are able to relate. Many people in positions of influence still consider Mr. Twain's definition of lies to be appropriate even today.

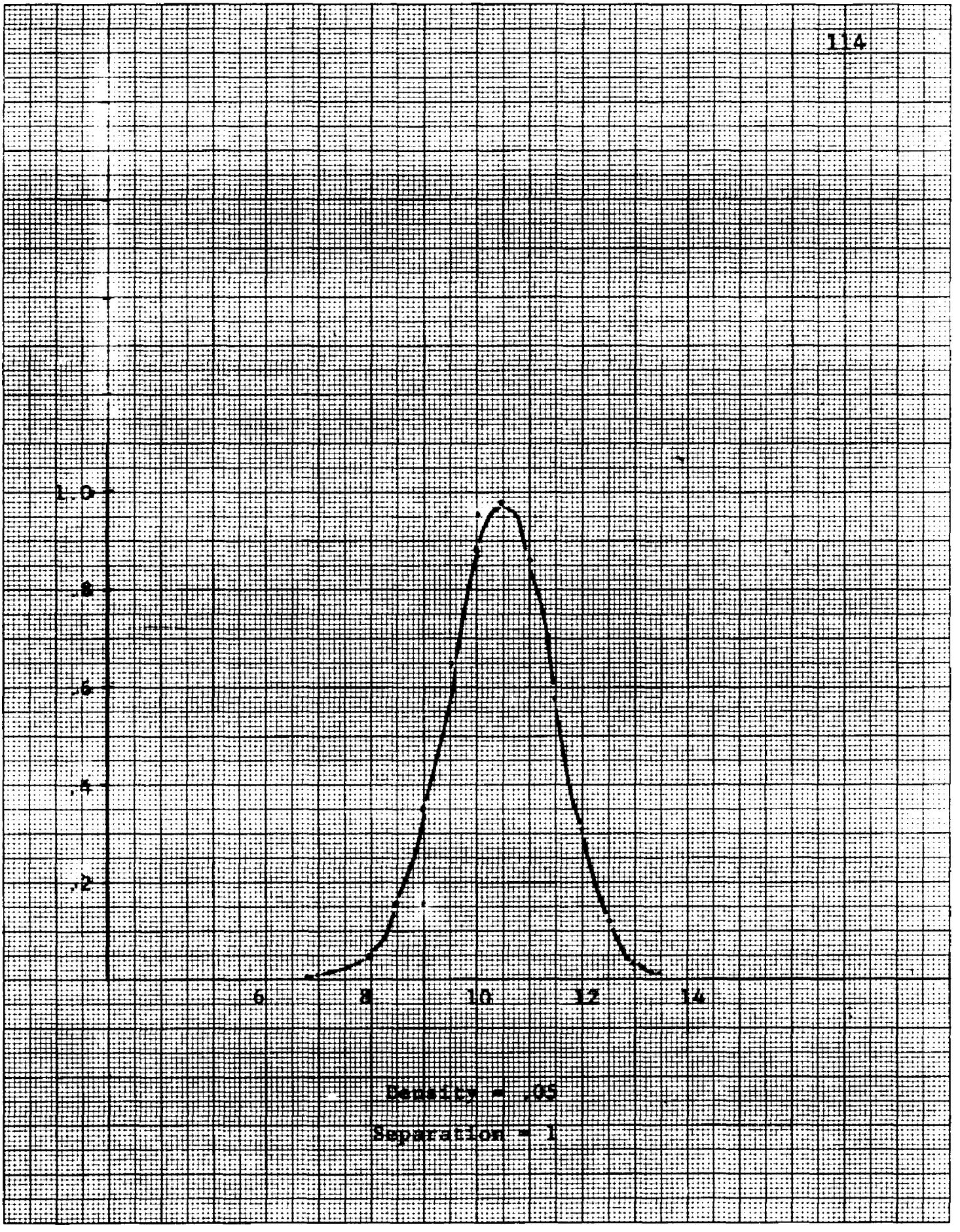
A statistical decision is based on probability, and there is always the possibility that the decision is an incorrect one. The best that a practitioner who is using statistics can hope to accomplish is a reduction in the probability of error. Statistics are not an extension of "damn lies". They are an attempt to measure what otherwise could not be measured. Those who use statistics and even more so those who interpret these statistics must not claim to have found the ultimate answer. A decision based on Statistic A may be directly opposed to the decision based on Statistic B. These apparent contradictions are especially true in the behavioral and social sciences where human beings cannot be measured with the precision of an experiment conducted in a laboratory setting with maximum control. Statistics only become "damn lies" when they are cloaked with an aura of absolute certainty.

The present study gives no definite answer to the question of Relative Power. A discussion of power requires the impossible, an exact knowledge of population shape. One can only use the model that appears to be most correct based

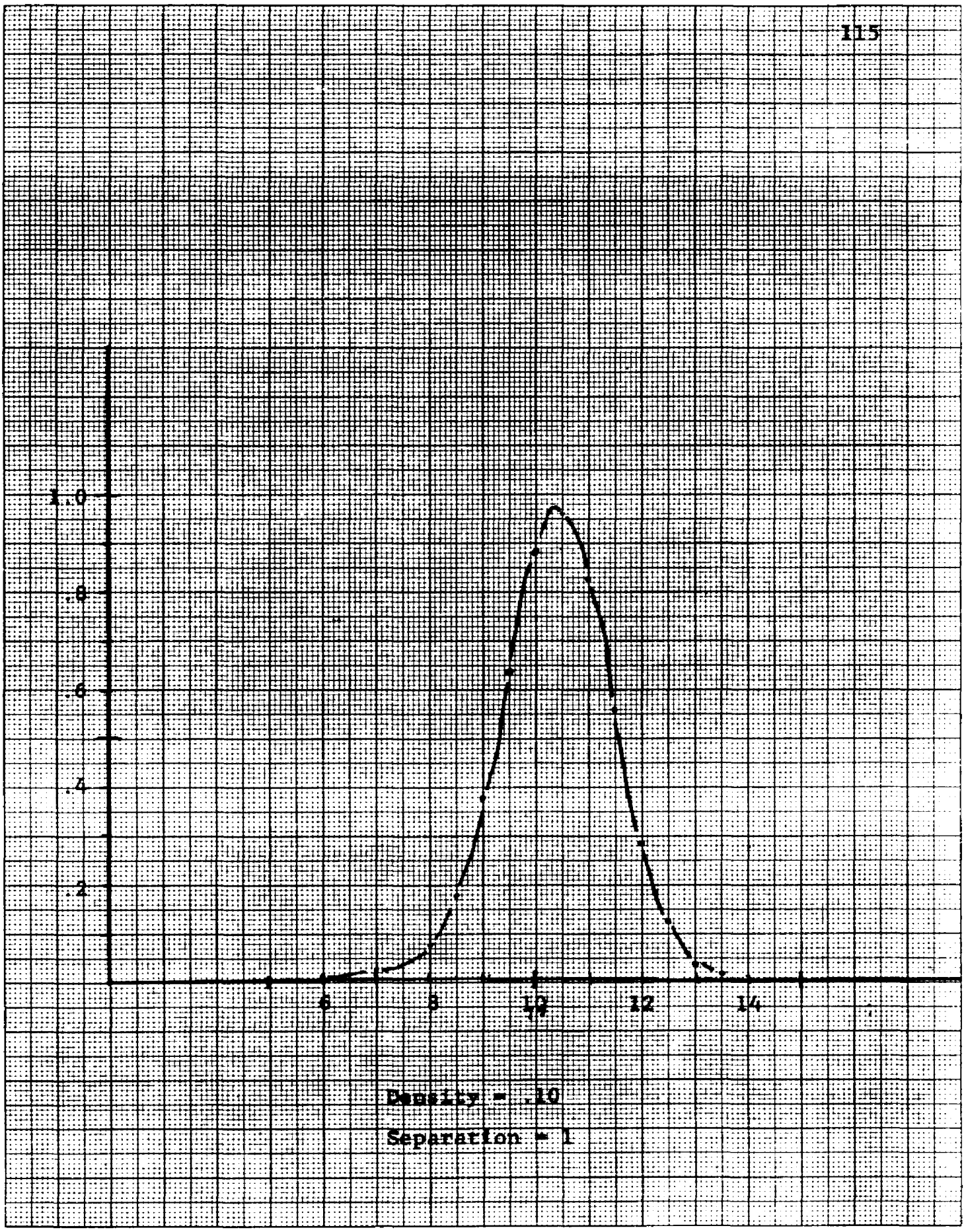
on the information that is available and the limitations of a given study. The choice between statistical tests is the choice between models. Hopefully, a researcher will choose the test that is least likely to err. One is never certain. Only with much further investigation can the question of Relative Power between the "Student" t-test and the Mann-Whitney U-Test be brought into a clearer light. As of now, the picture is opaque. As stated earlier, truth is not static. If society is willing to invest the energy to progress in small increments, eventually the darkness will begin to dissipate.

APPENDIX A

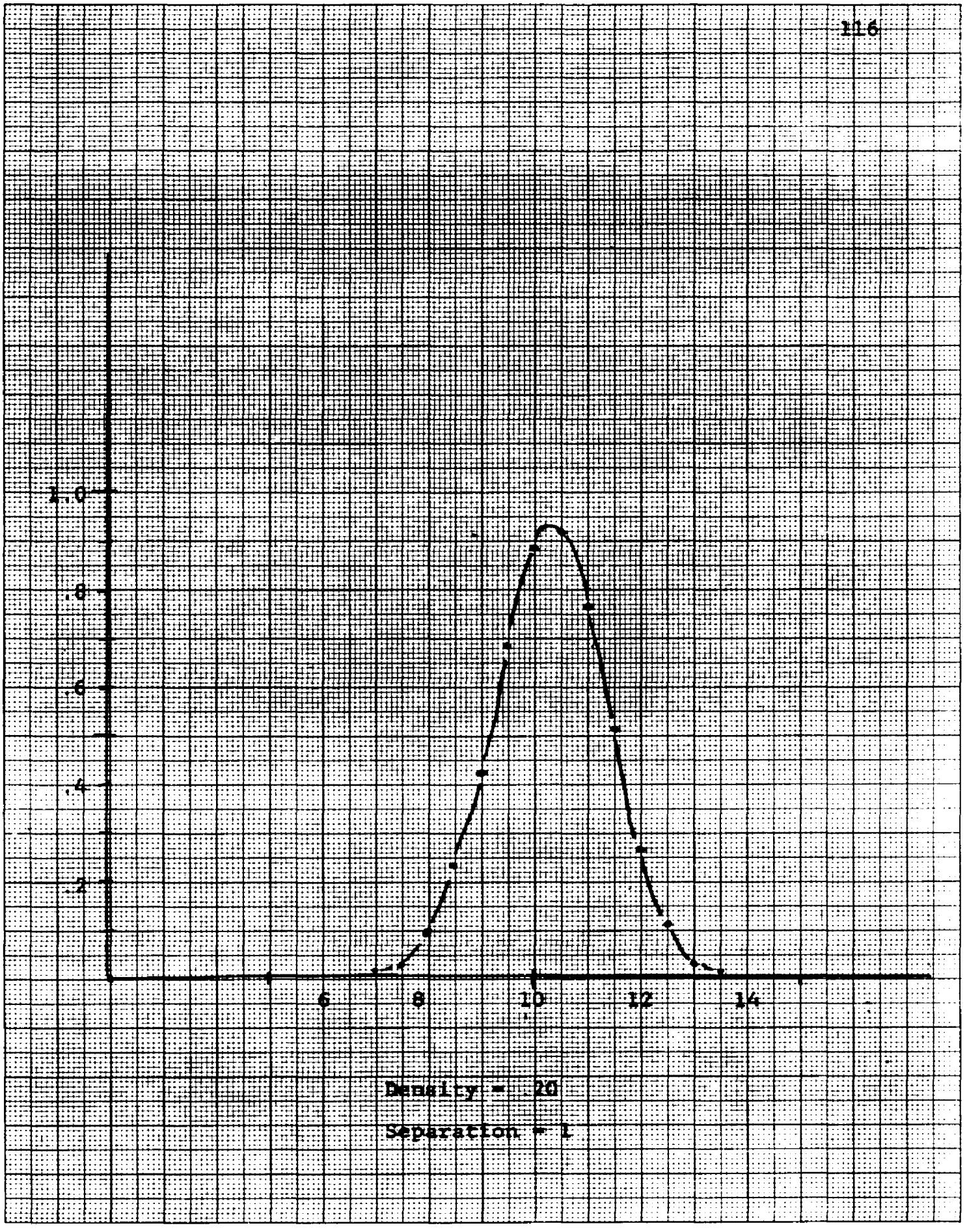
Graphs of Mixed-Normal Population Density Distribution



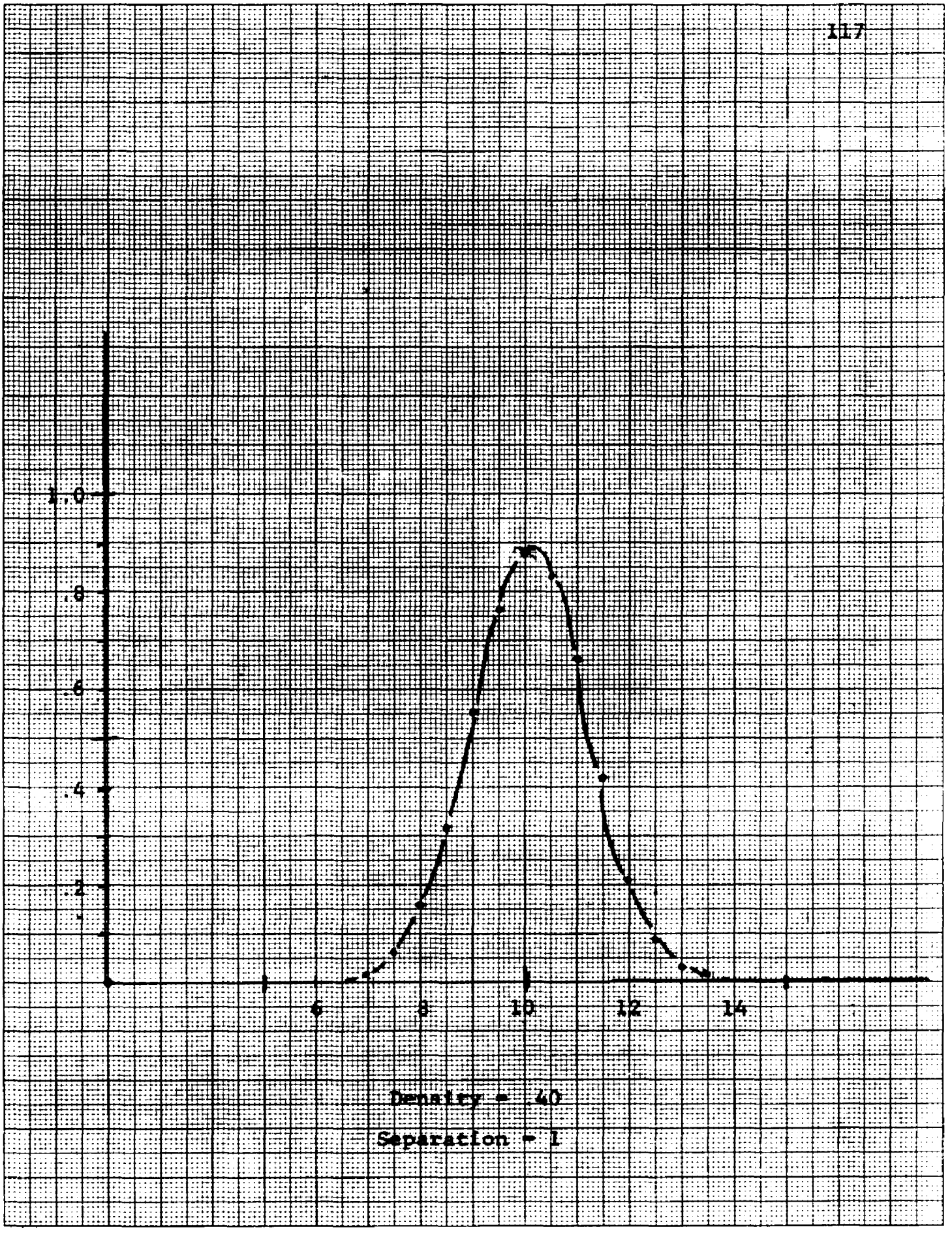
Millimeters to the Centimeter



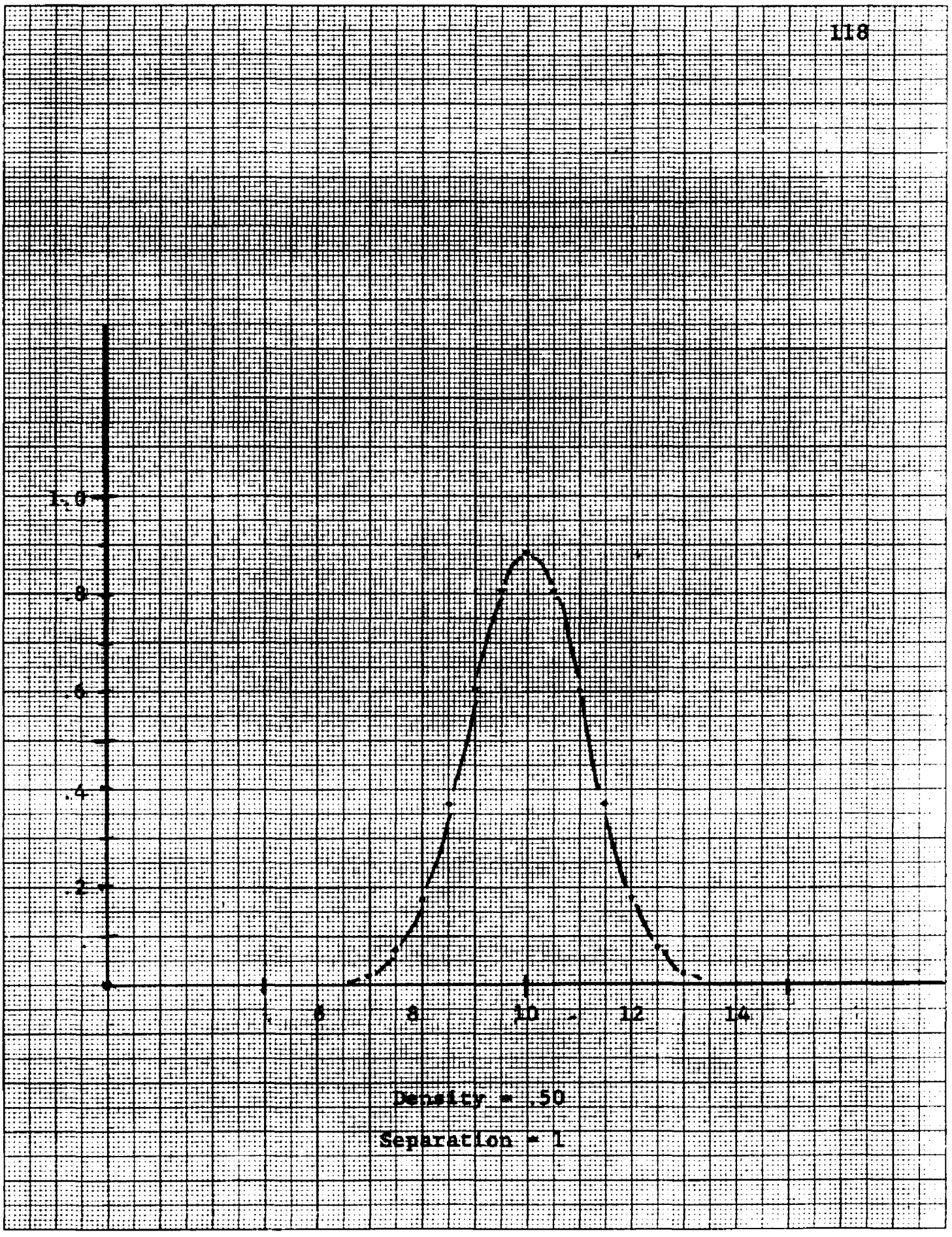
Millimeters to the Centimeter

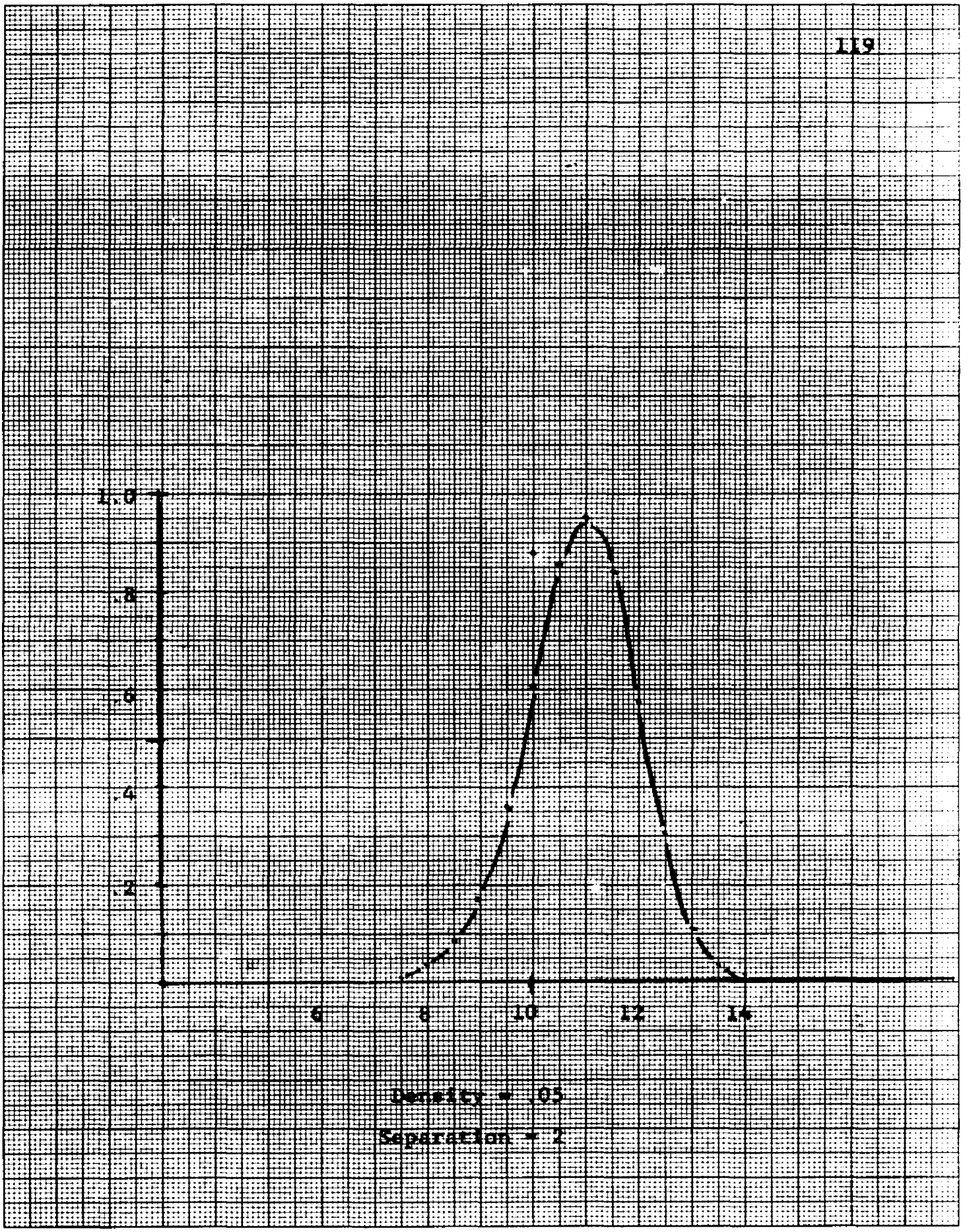


imeters to the Centimeter

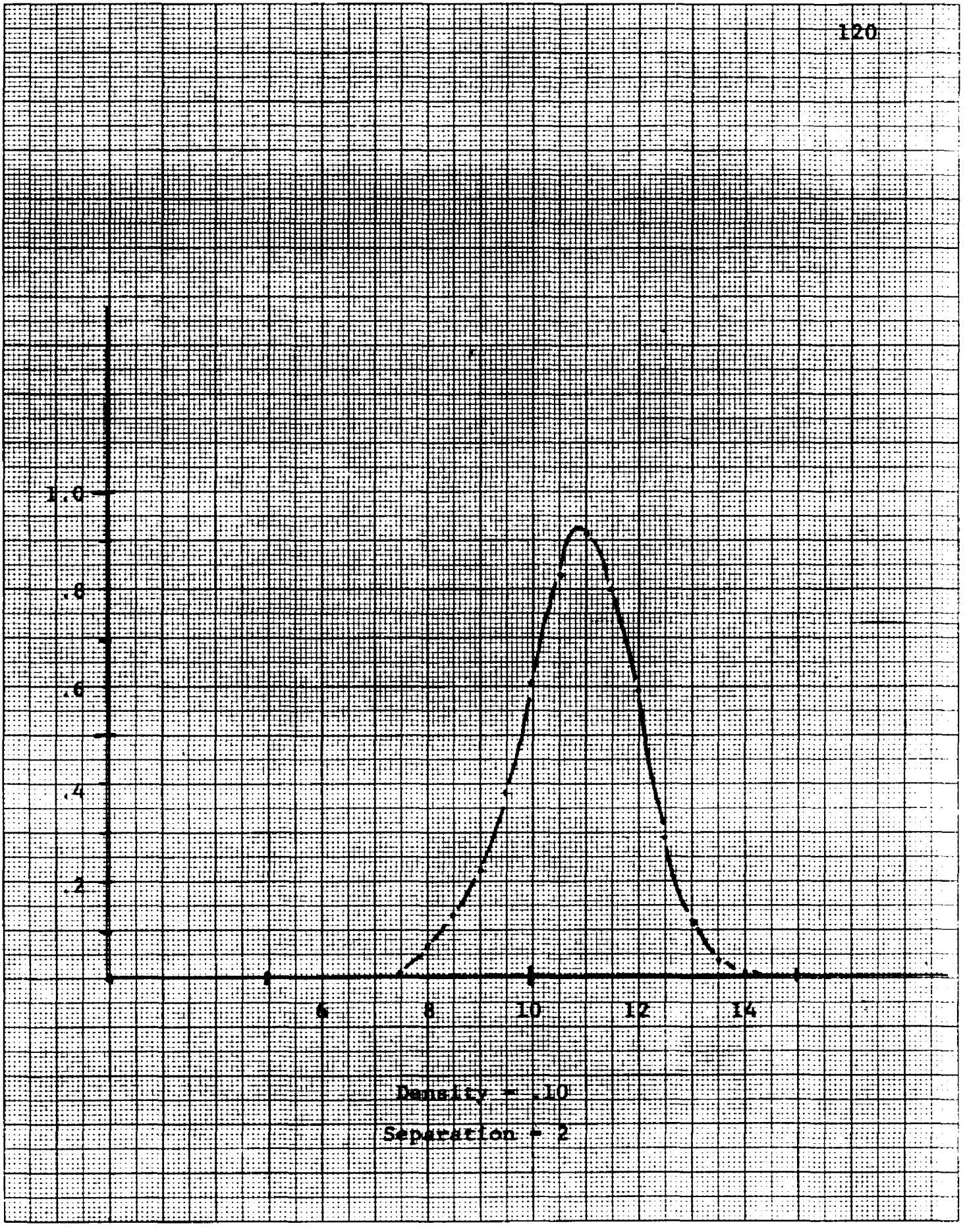


limeters to the Centimeter

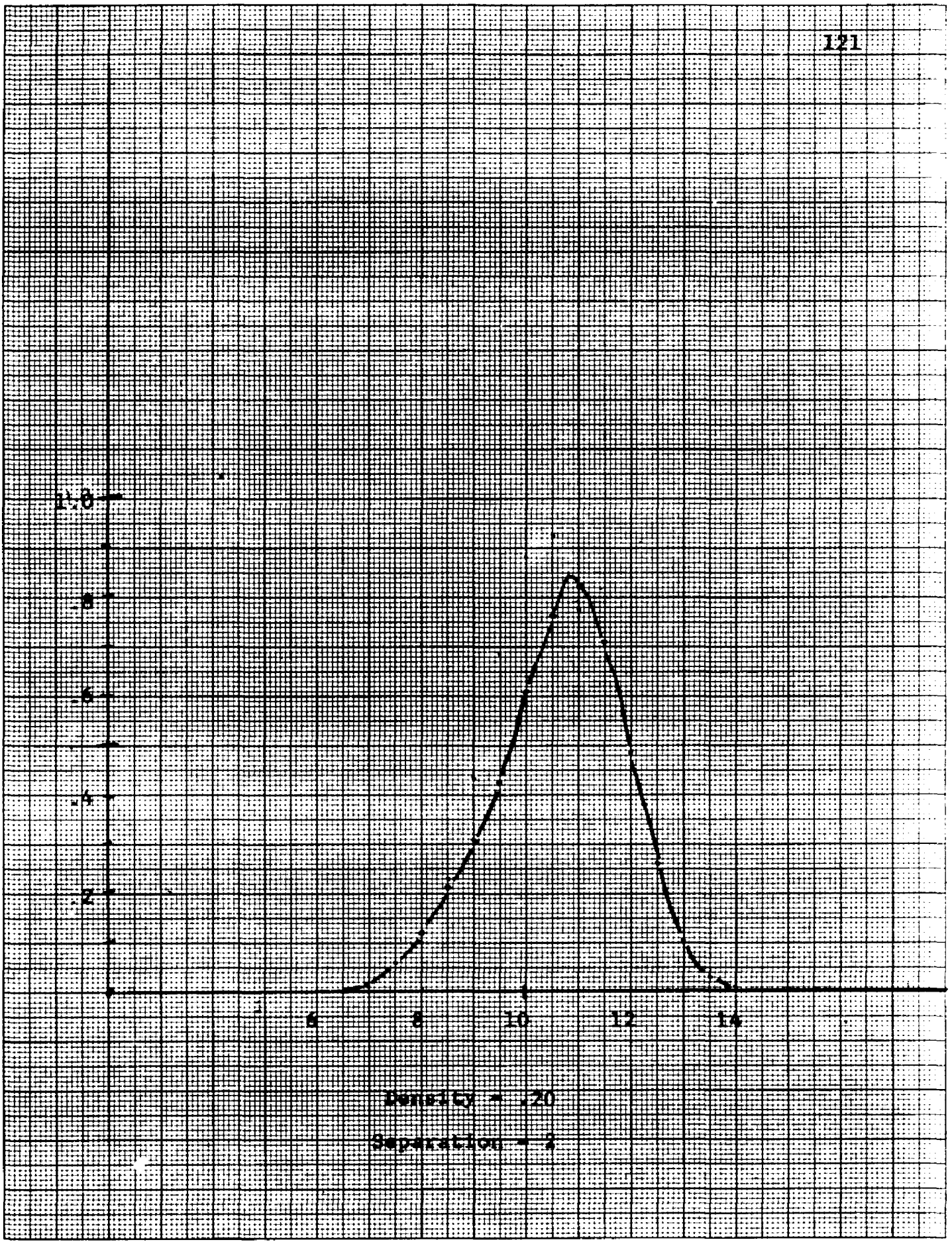

1.0
.8
.6
.4
.2
6
8
10
12
14
Density = .50
Separation = 1



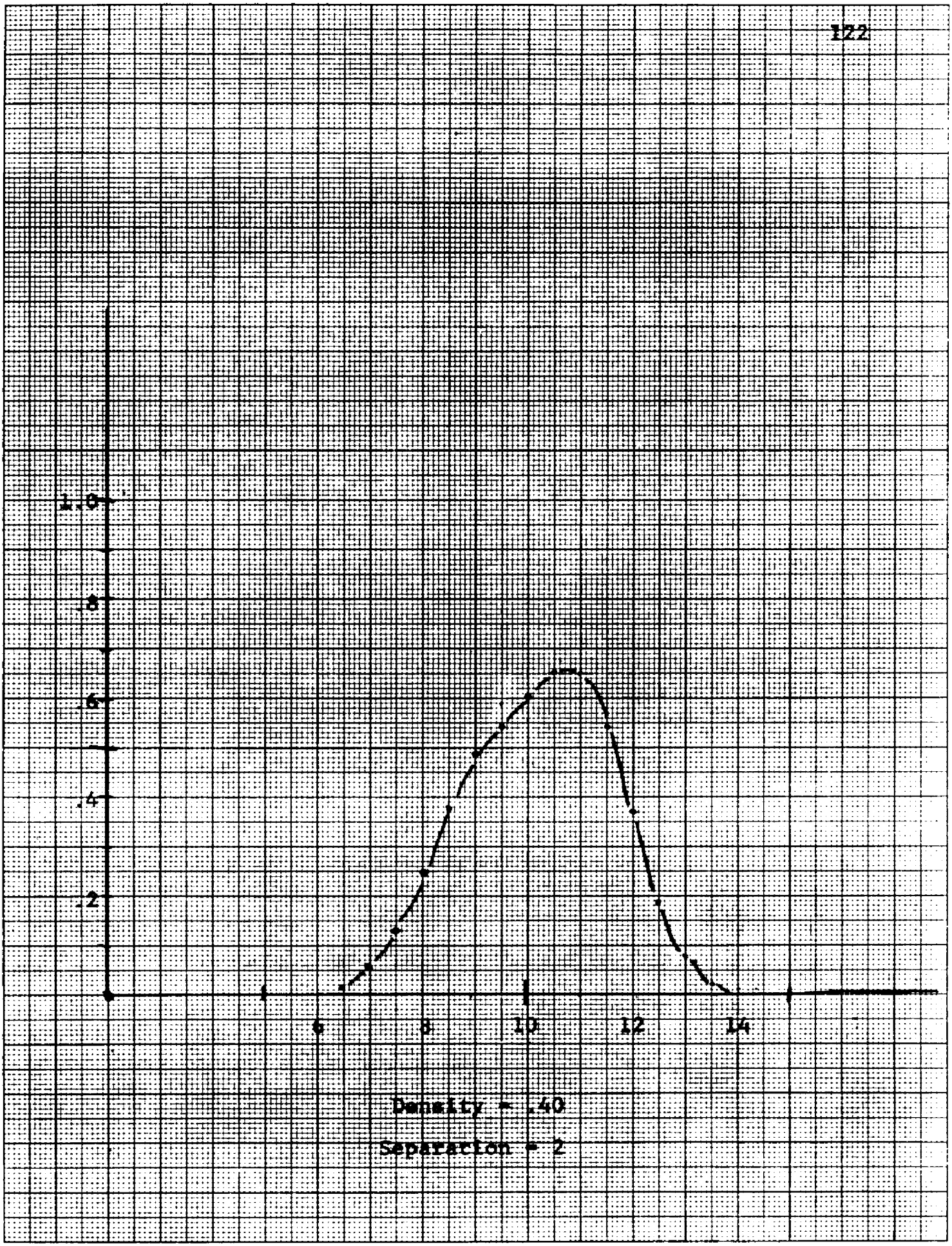
Millimeters to the Centimeter

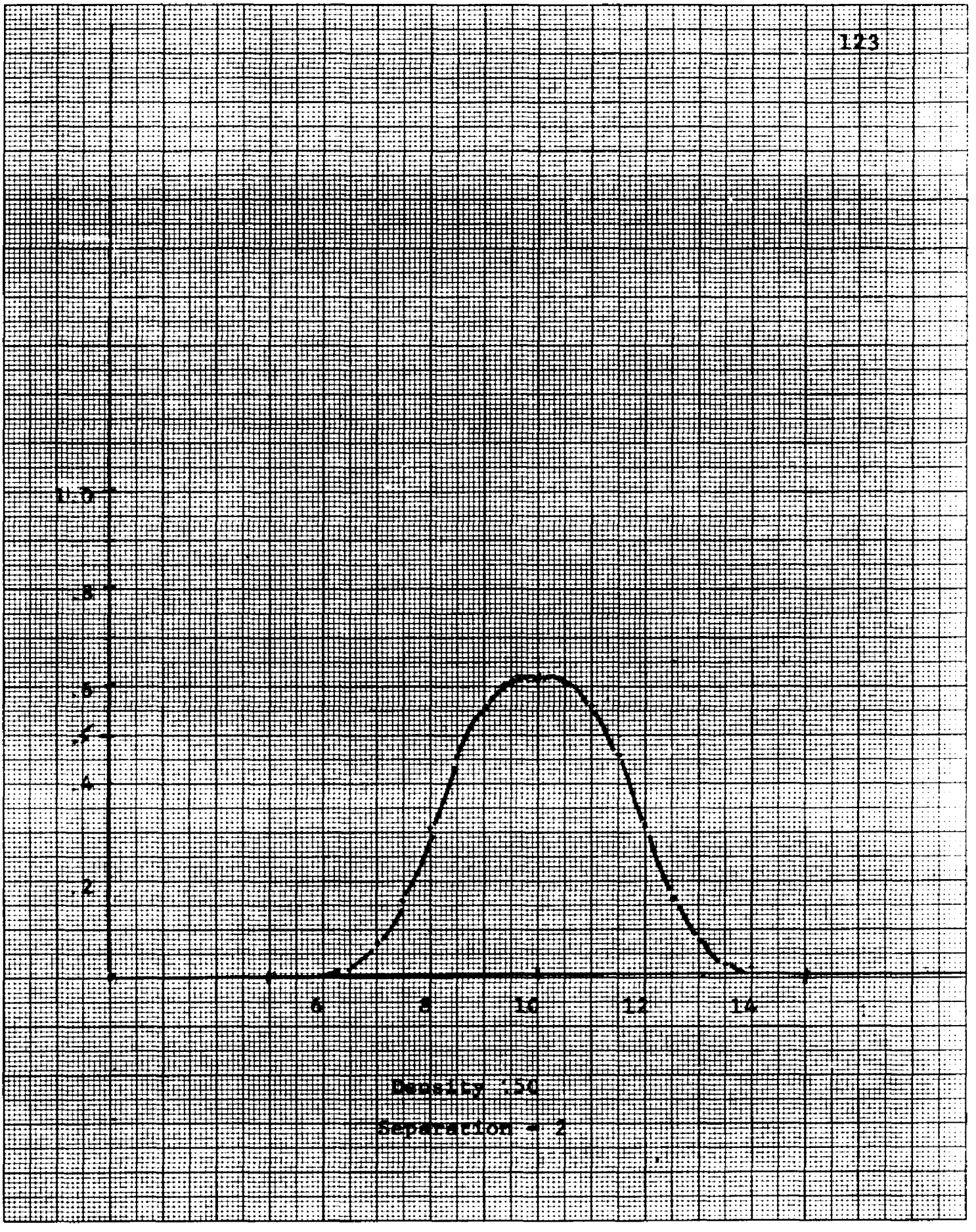


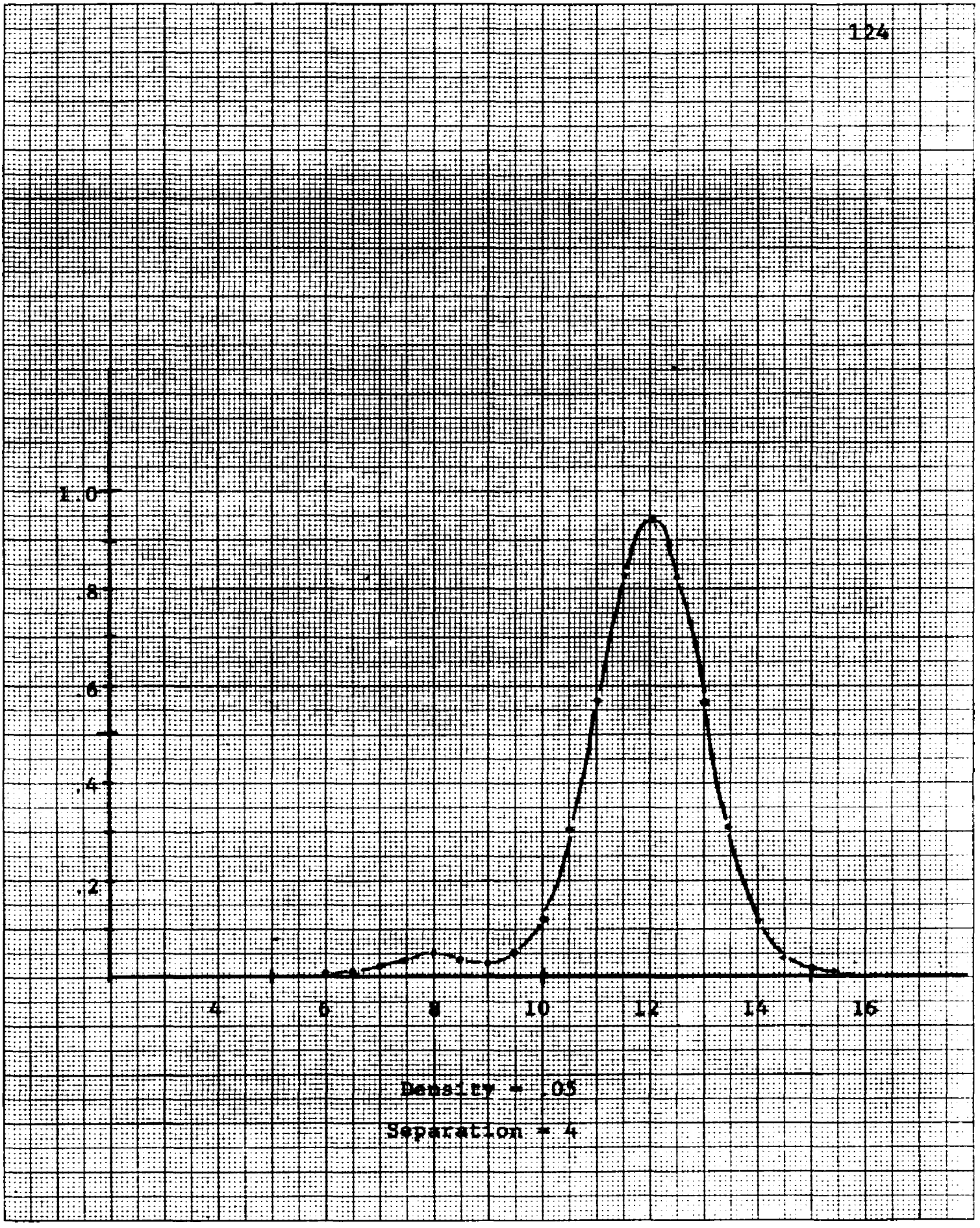
Millimeters to the Centimeter



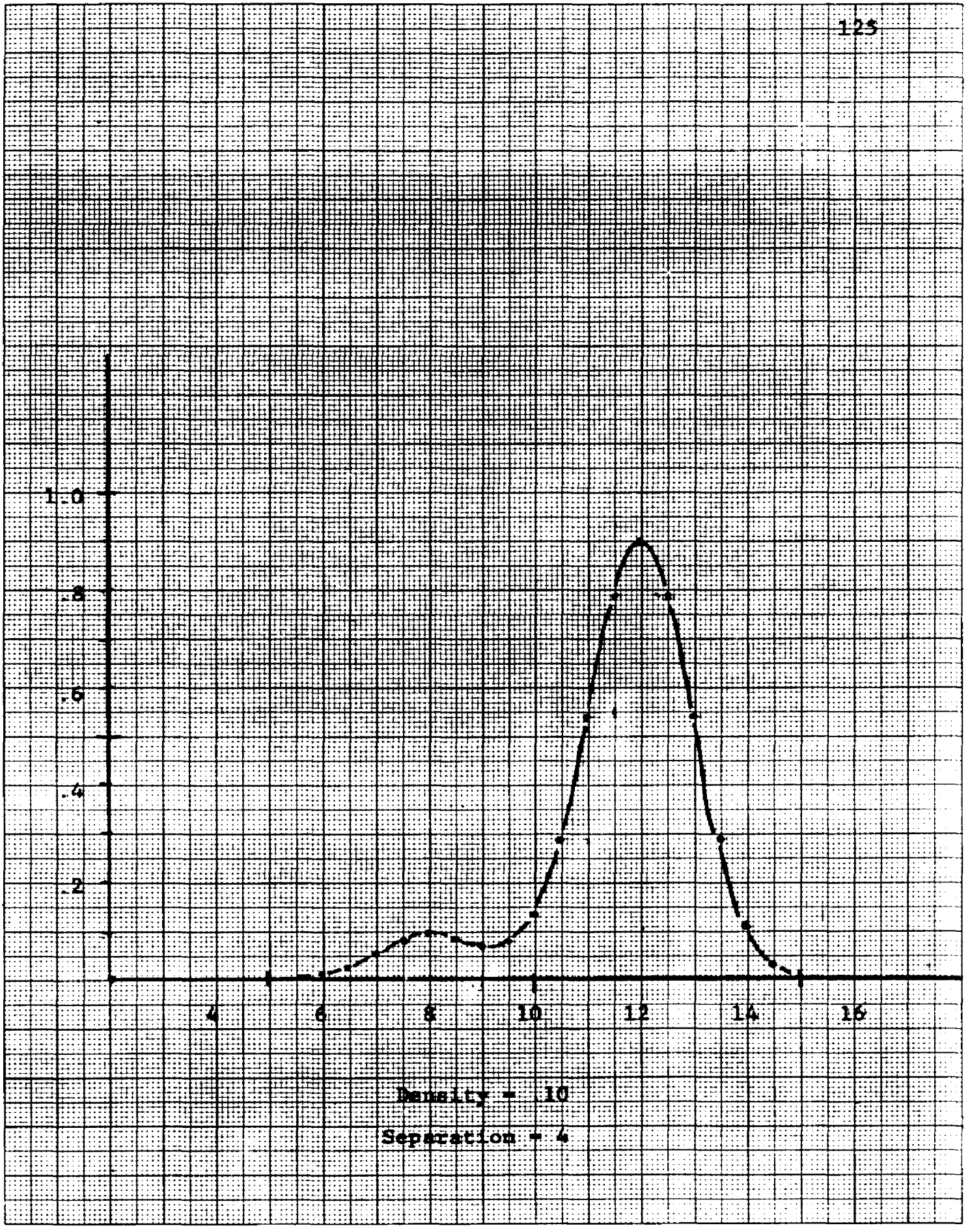
Millimeters to the Centimeter

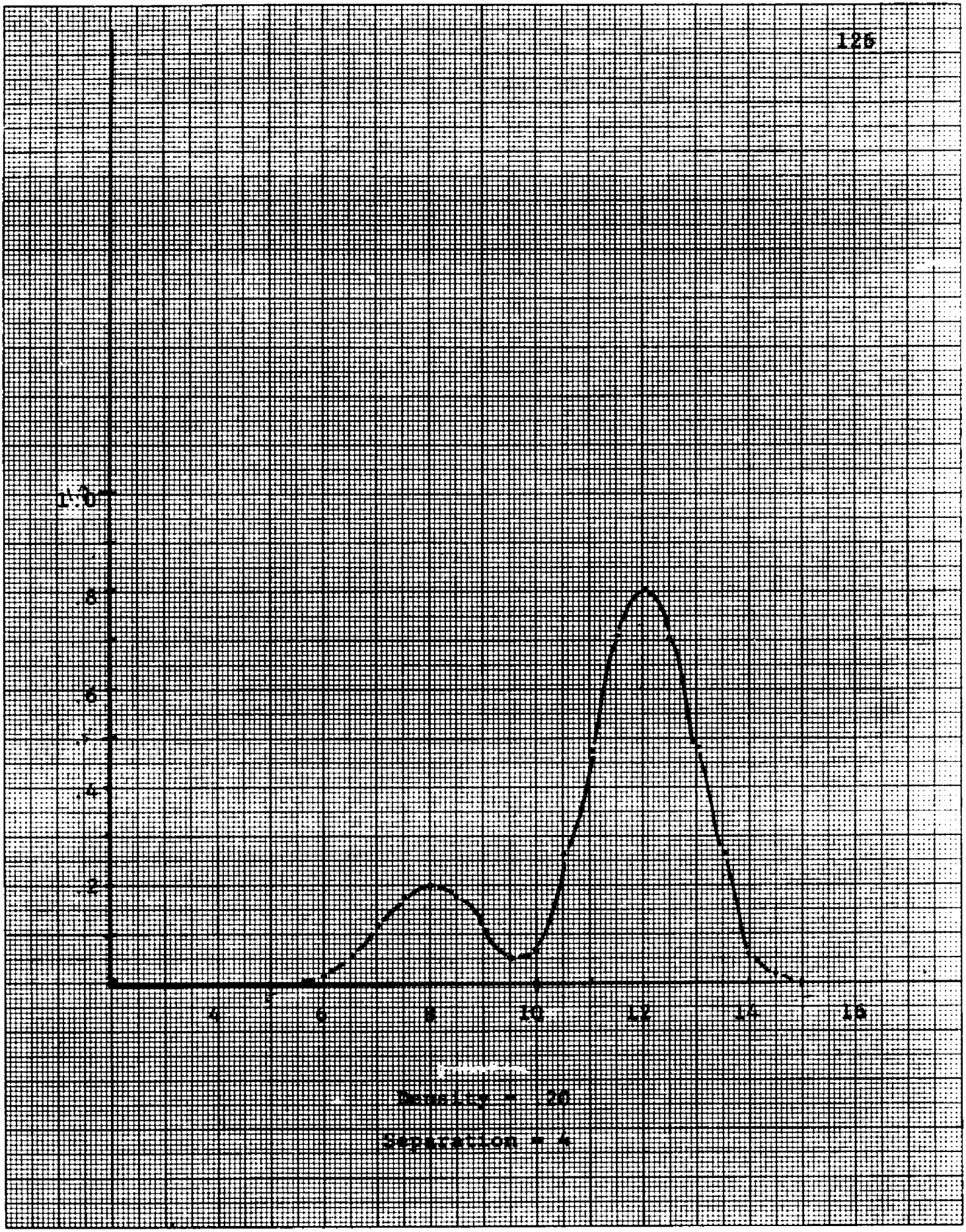

1.0
.8
.6
.4
.2
6
8
10
12
14
Density = .40
Separation = 2


1.0
.8
.6
.4
.2
6
8
10
12
14
Density :50
Separation = 2

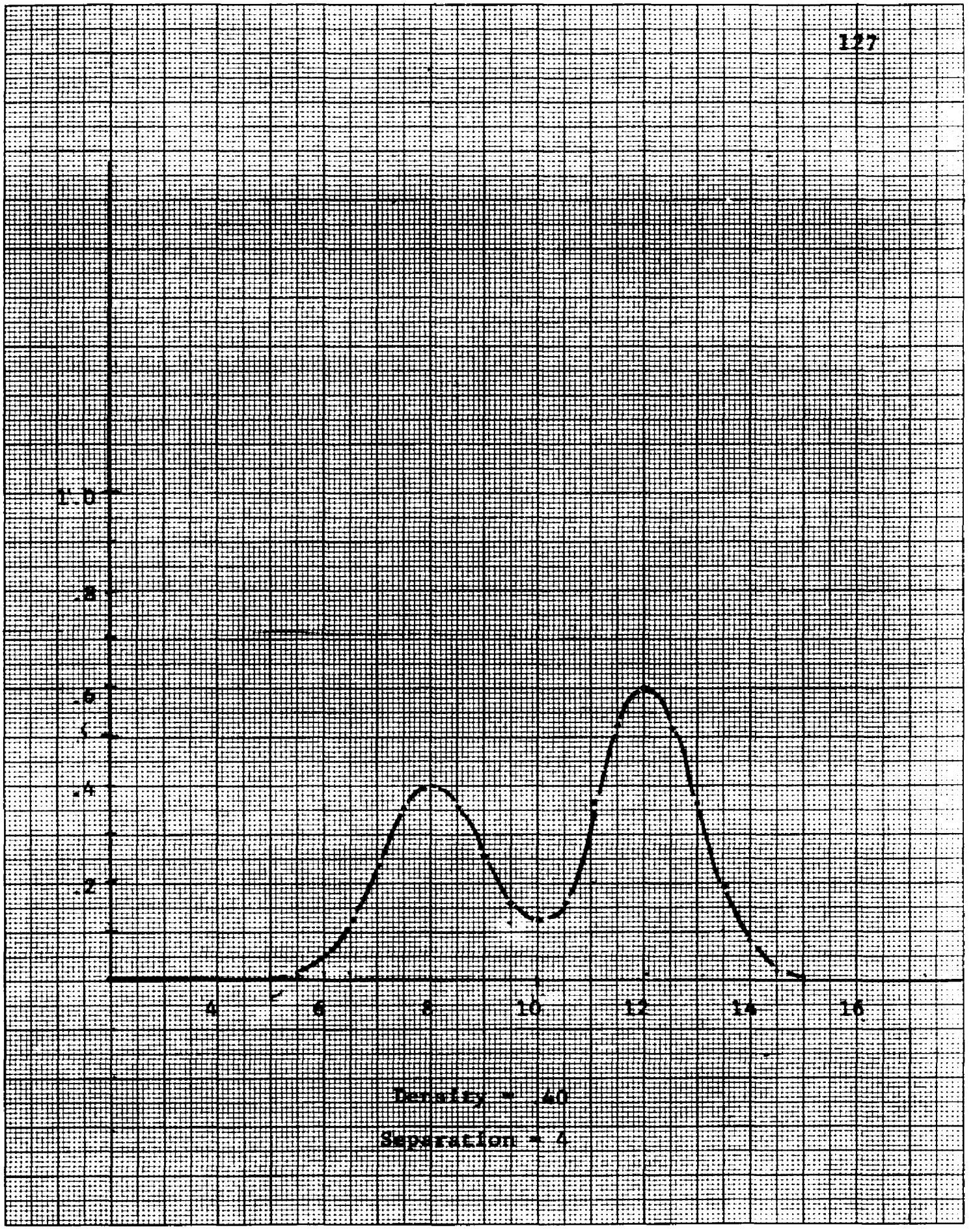


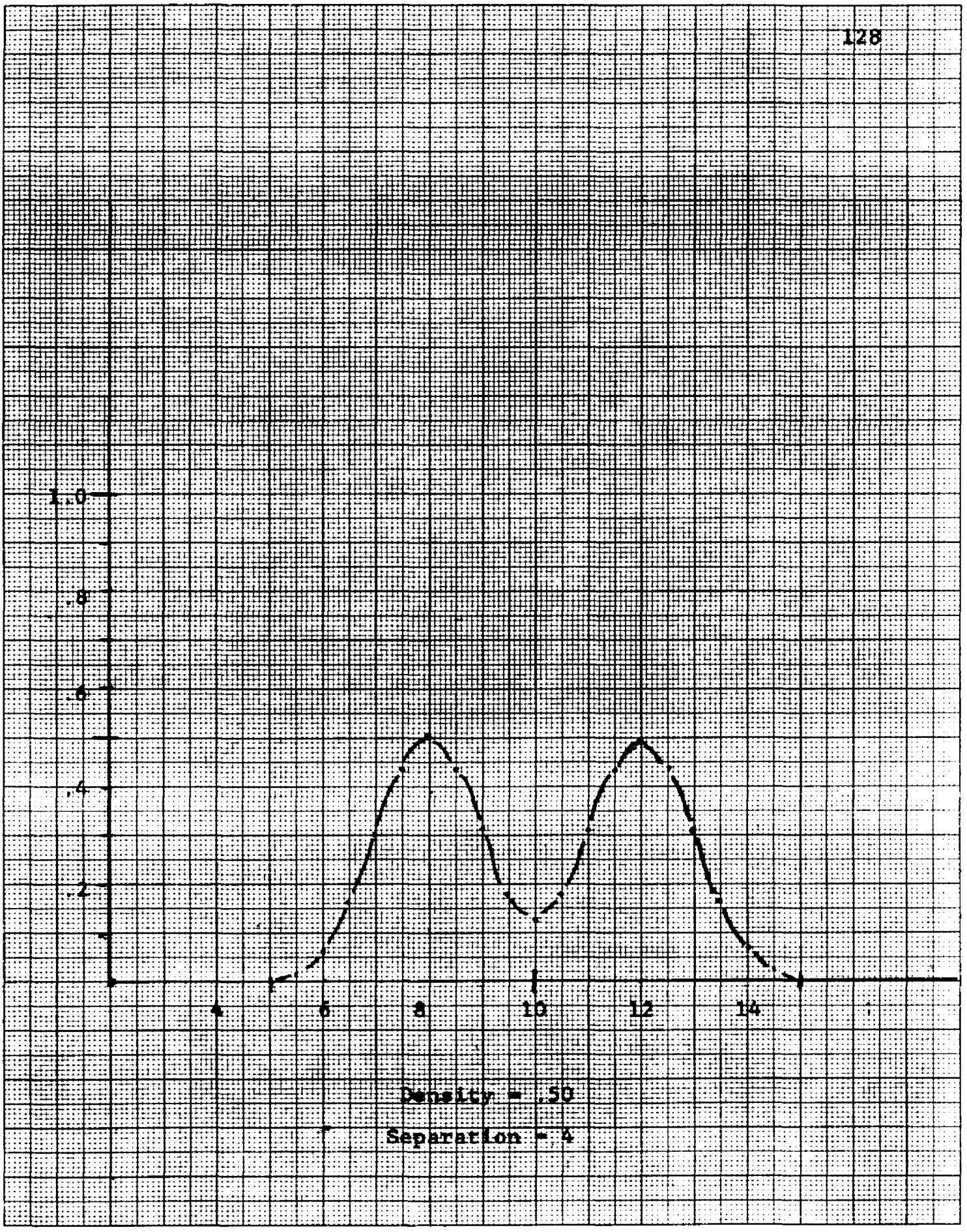
meters to the Centimeter


1.0
.8
.6
.4
.2
4
6
8
10
12
14
16
Density = .10
Separation = 4

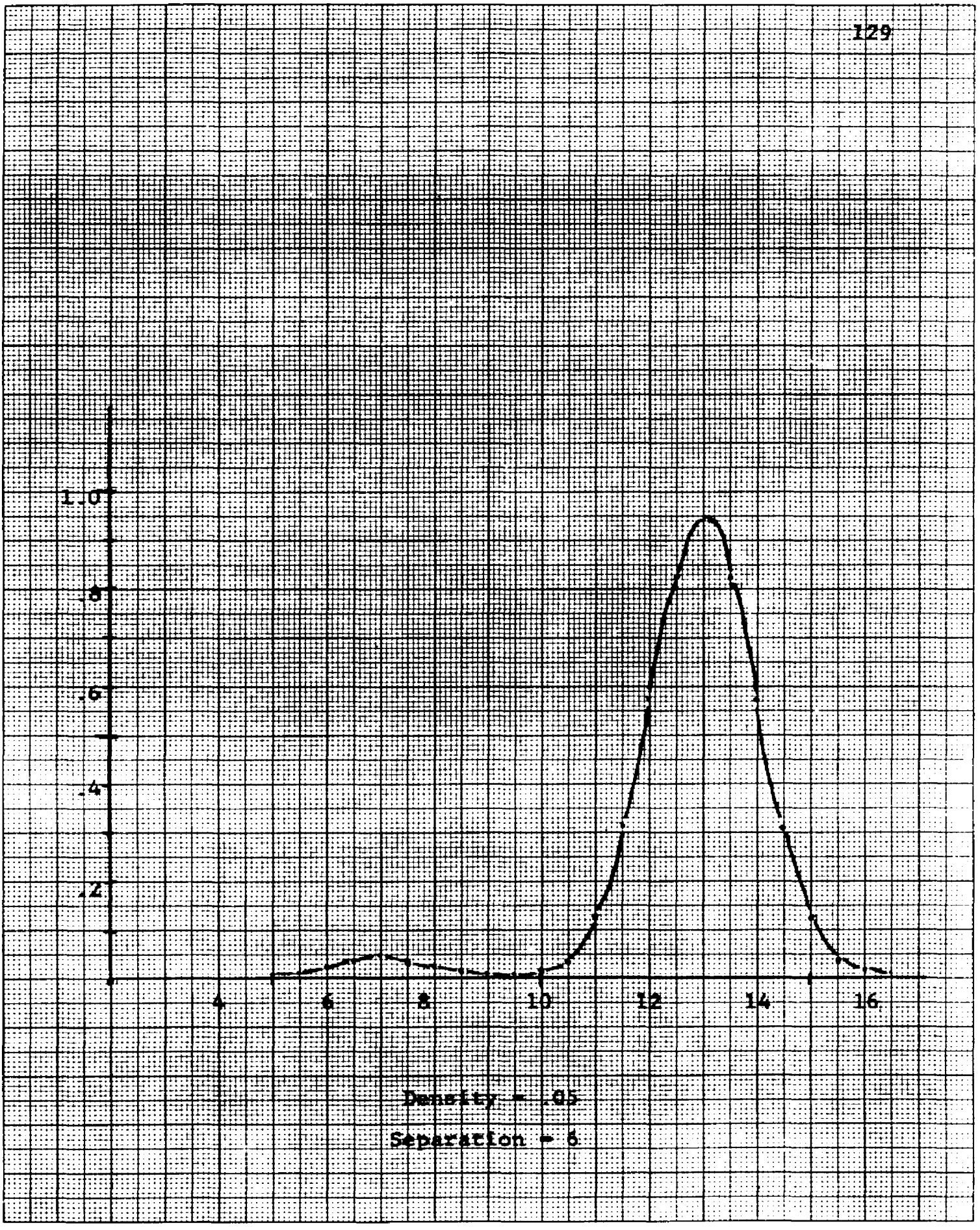


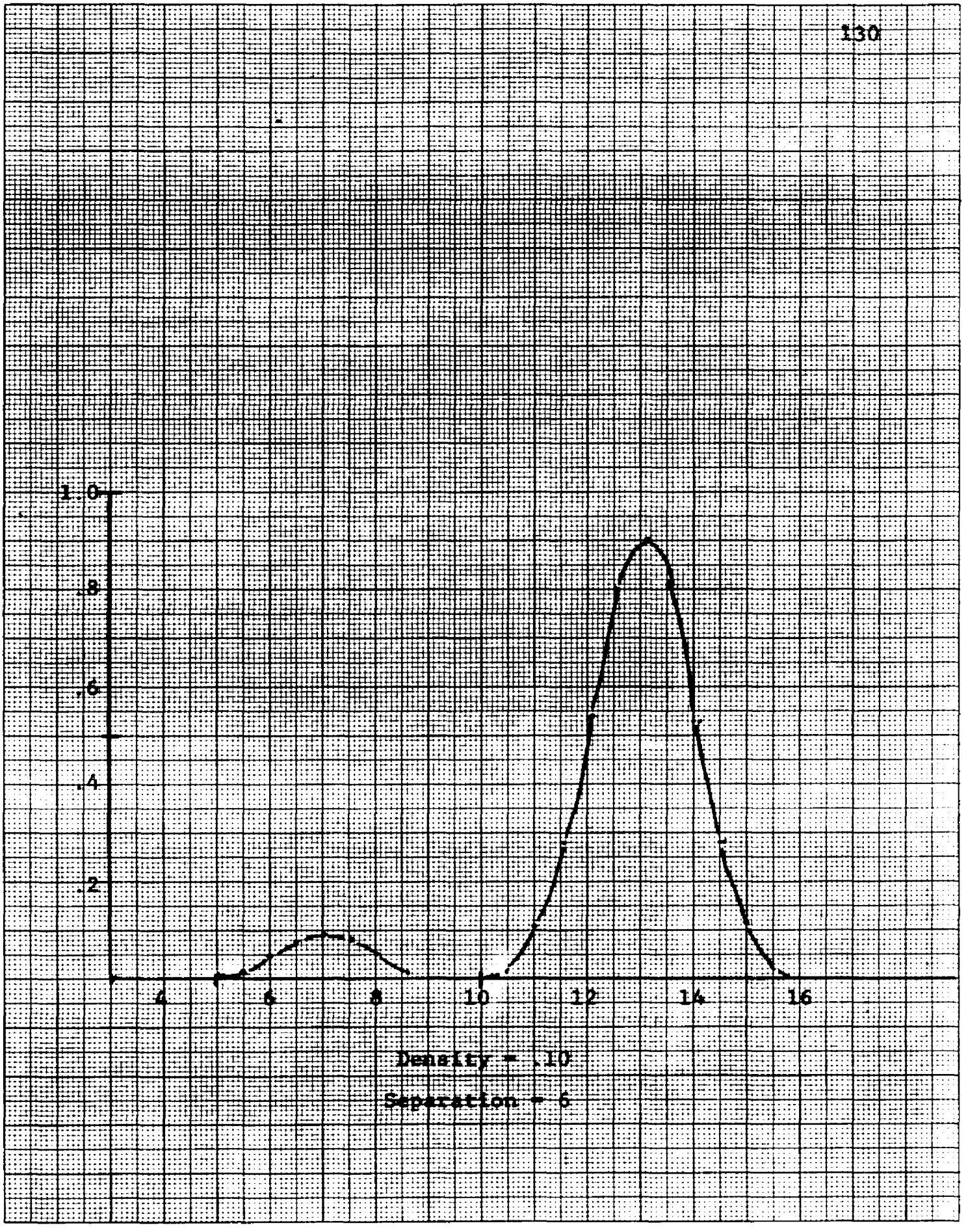
Millimeters to the Centimeter


1.0
.8
.6
.5
.4
.2
4
6
8
10
12
14
16
Density = .40
Separation = 4

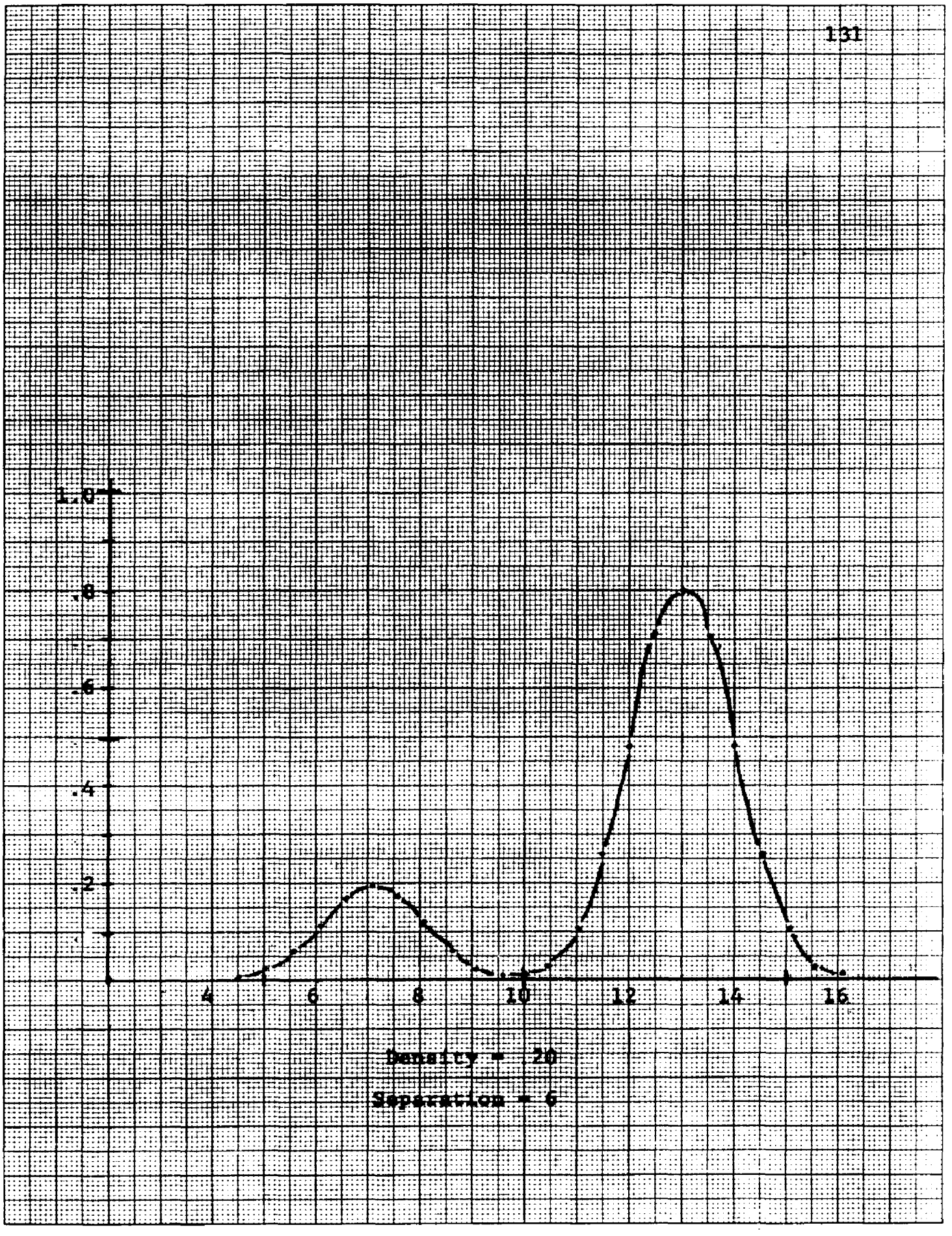


meters to the Centimeter

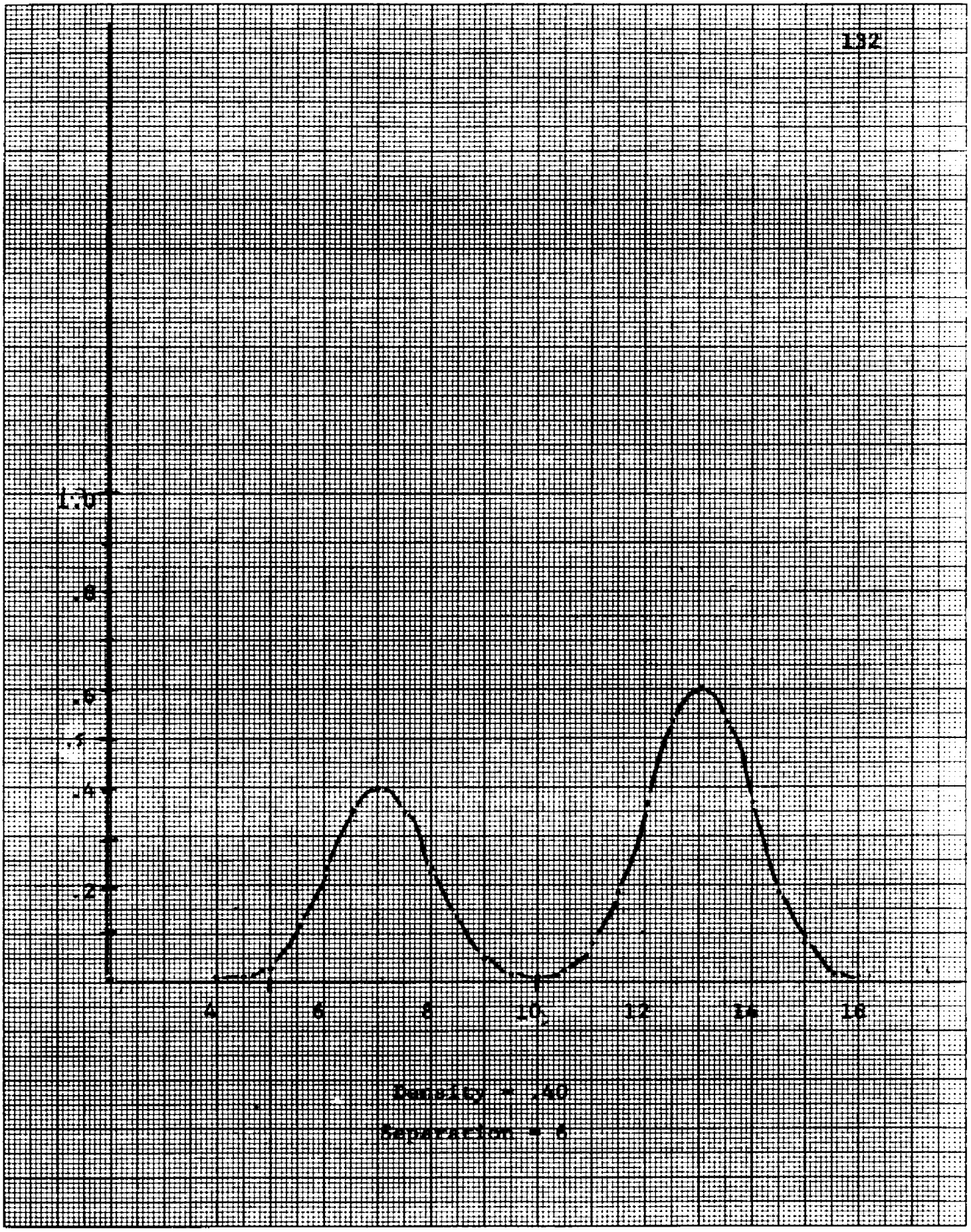

1.0
.8
.6
.4
.2
4
6
8
10
12
14
16
Density = .05
Separation = 6

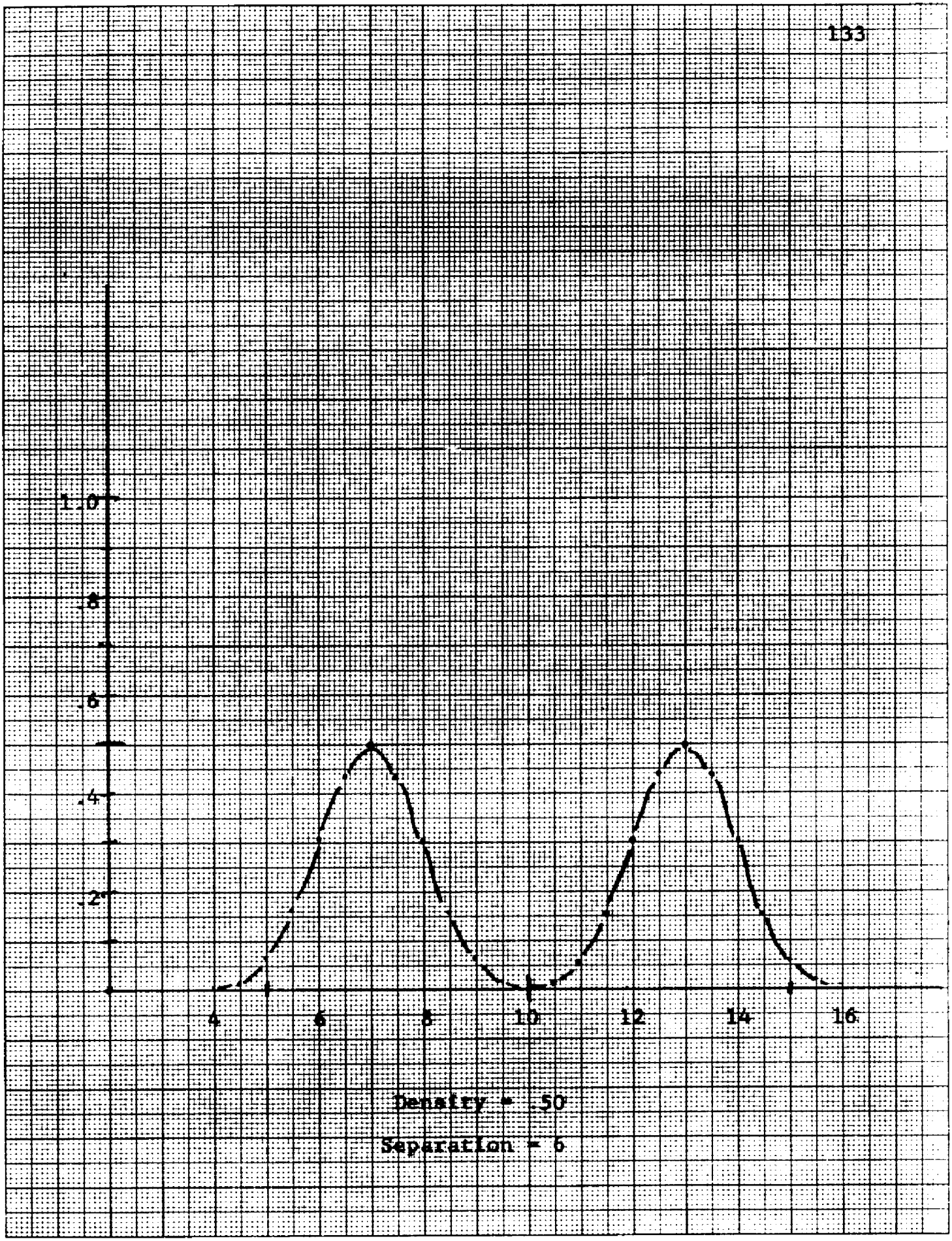


meters to the Centimeter

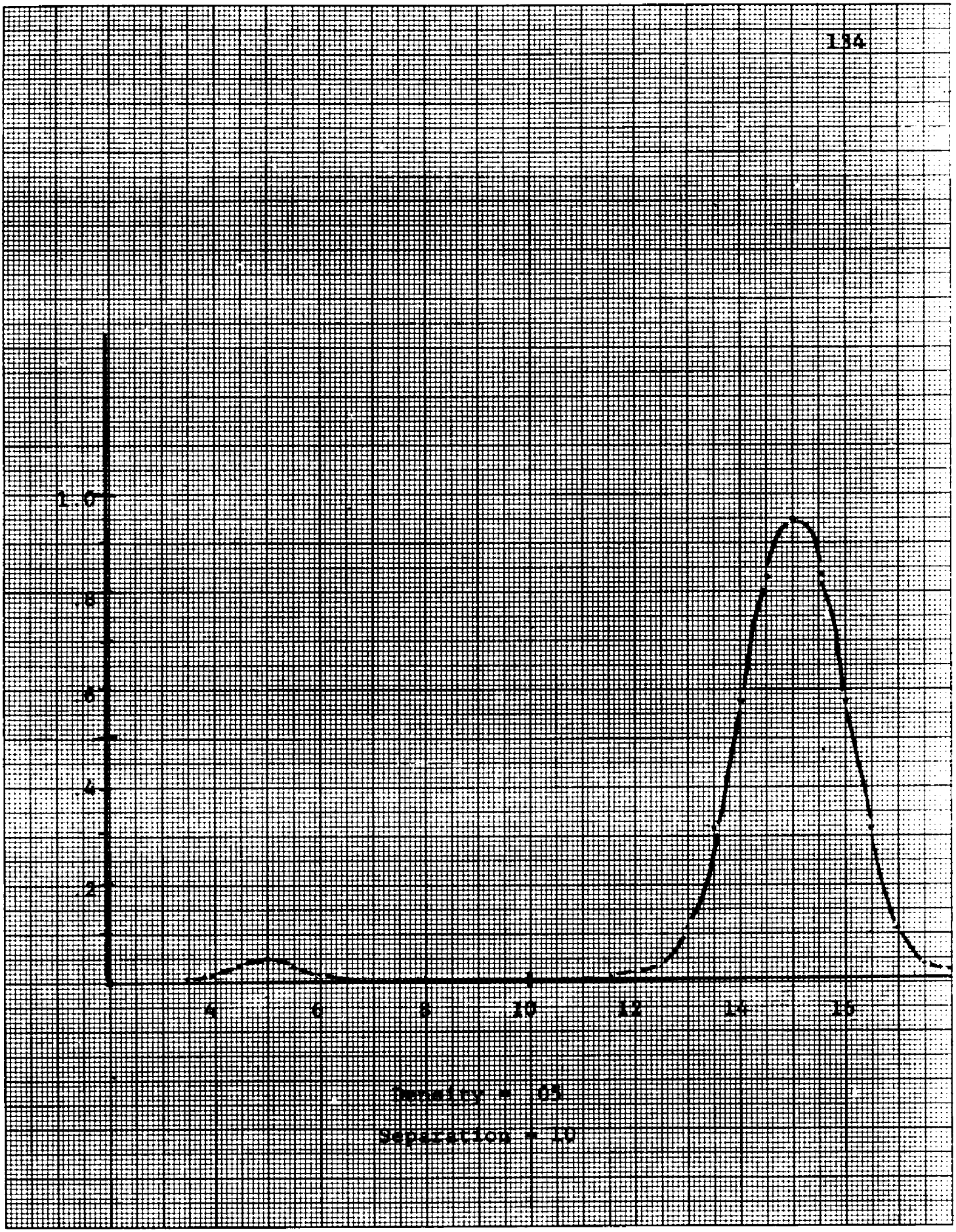


Millimeters to the Centimeter

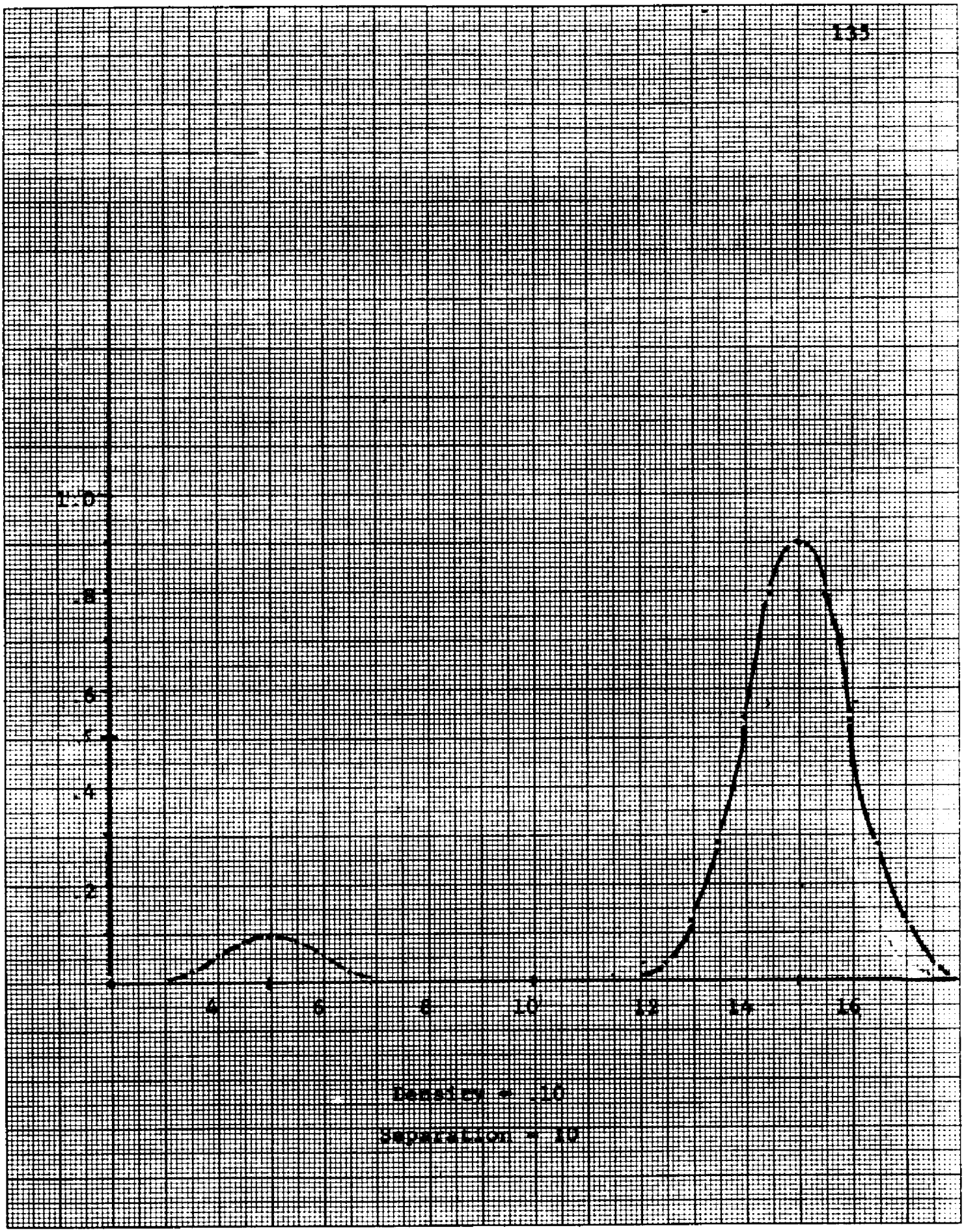

1.0
.8
.6
.4
.2
4
6
8
10
12
14
16
Density = .40
Separation = 6

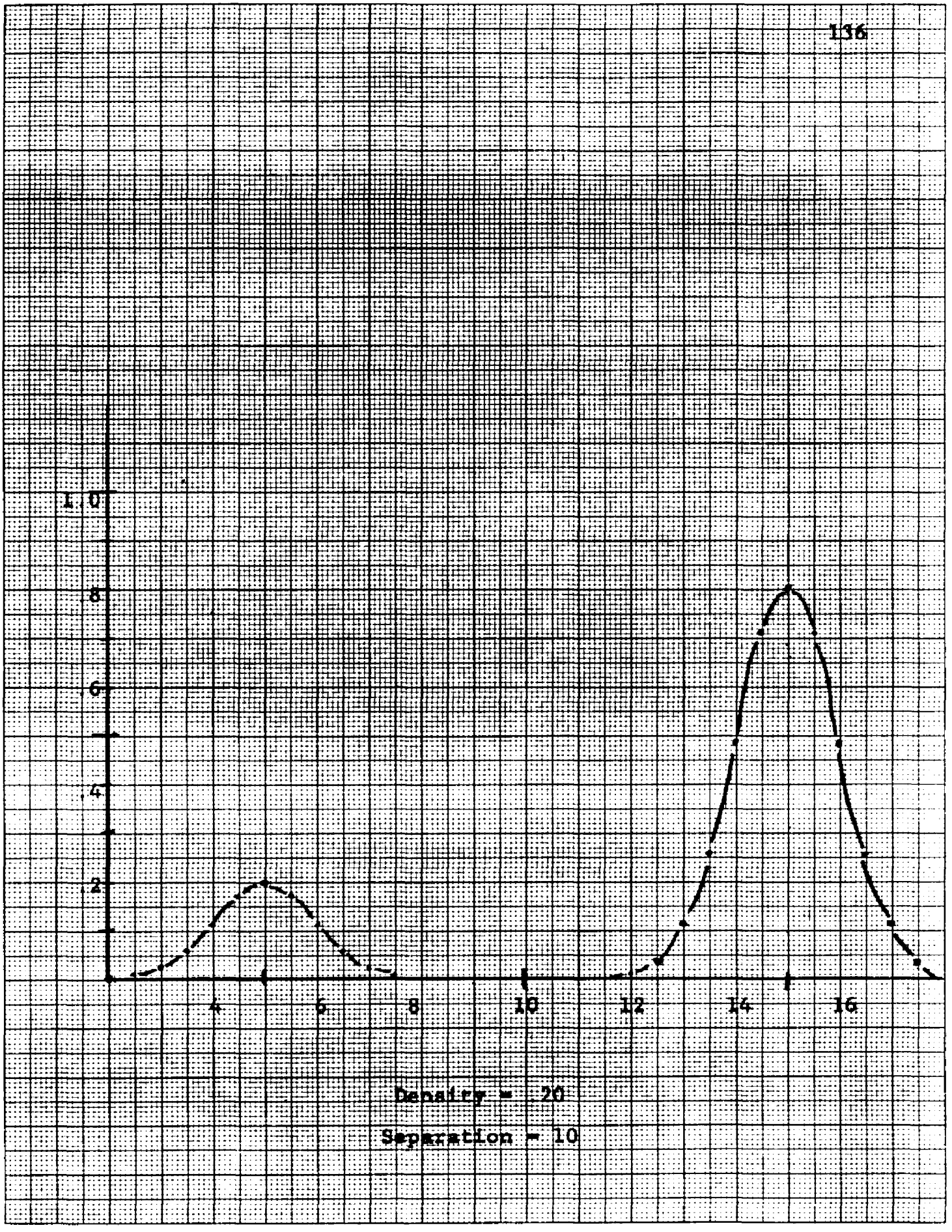


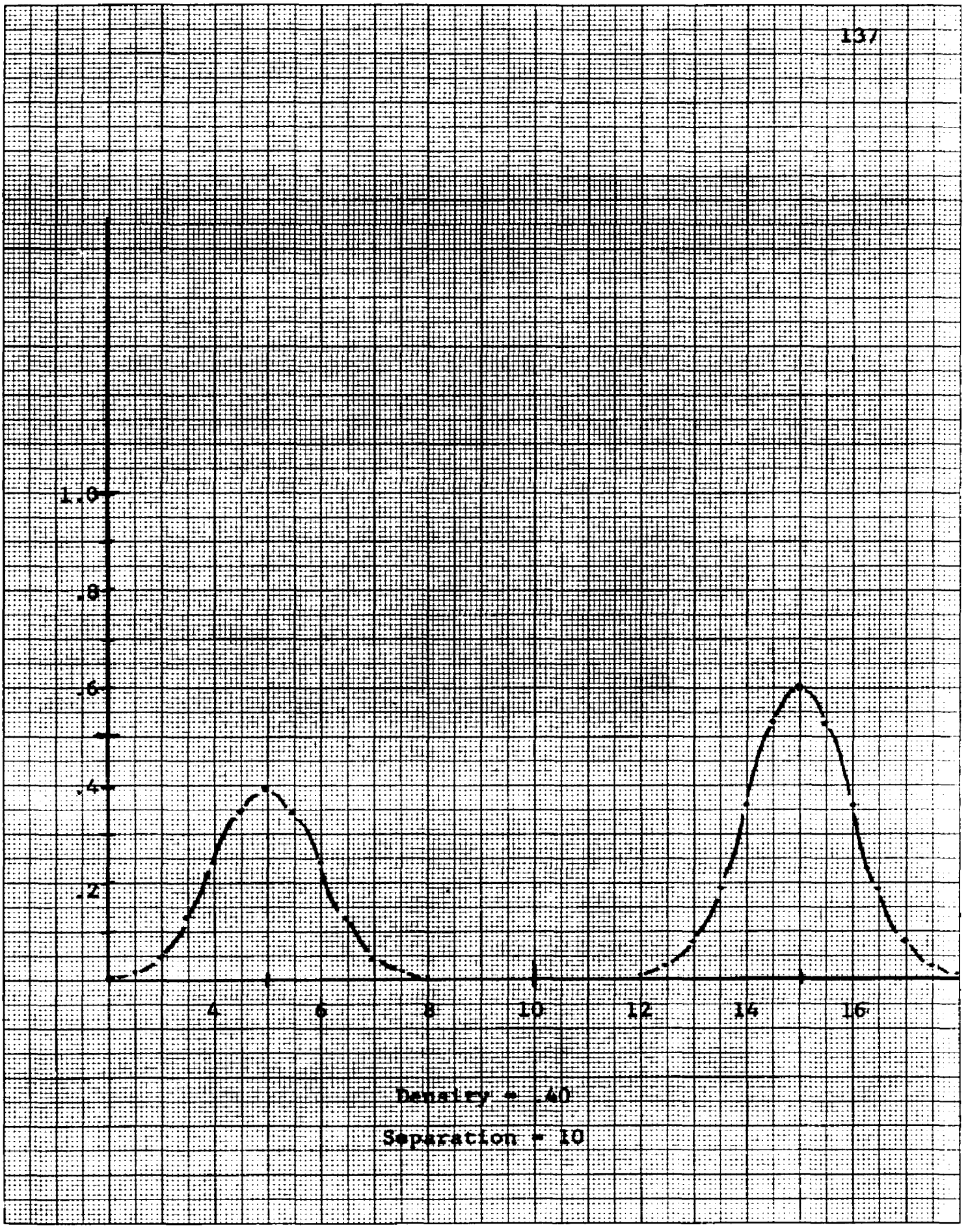
illimeters to the Centimeter



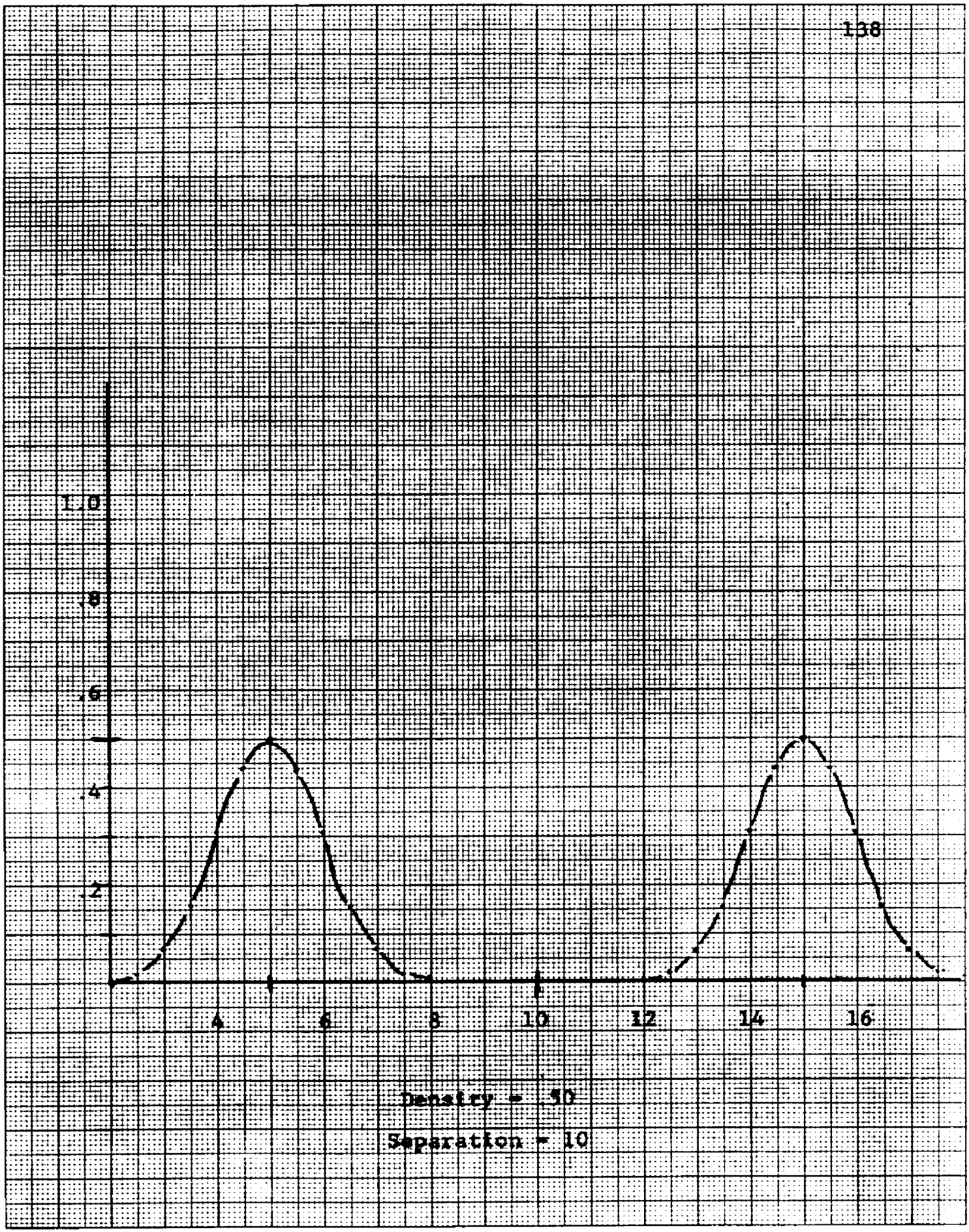
Millimeters to the Centimeter


1.0
.8
.6
.4
.2
4
6
8
10
12
14
16
Density = .10
Separation = 10


1.0
.8
.6
.4
.2
4
6
8
10
12
14
16
Density = .20
Separation = 10



Rmeters to the Centimeter



meters to the Centimeter

APPENDIX B

Means, Standard Deviations, and Coordinates for Mixed-Normal Density Distributions

Density = .05

Separation = 1

x	f(x)
6	.000
6.5	.001
7	.004
7.5	.017
8	.058
8.5	.159
9	.353
9.5	.626
10	.882
10.5	.980
11	.855
11.5	.583
12	.311
12.5	.129
13	.042
13.5	.011
14	.002
14.5	.000

Mean = 10.450

Standard Deviation = 1.001

Density = .05

Separation = 2

x	f(x)
6	.000
6.5	.002
7	.007
7.5	.018
8	.041
8.5	.086
9	.179
9.5	.353
10	.607
10.5	.855
11	.957
11.5	.841
12	.577
12.5	.309
13	.129
13.5	.042
14	.011
14.5	.002
15	.000

Mean = 10.900

Standard Deviation = 1.005

Density = .05

Separation = 4

x	f(x)
5	.001
5.5	.002
6	.007
6.5	.016
7	.030
7.5	.044
8	.050
8.5	.046
9	.041
9.5	.058
10	.135
10.5	.311
11	.577
11.5	.838
12	.950
12.5	.838
13	.576
13.5	.308
14	.129
14.5	.042
15	.011
15.5	.002
16	.000

Mean = 11.800

Standard Deviation = 1.020

Density = .05

Separation = 6

x	f(x)
4	.001
4.5	.002
5	.007
5.5	.016
6	.030
6.5	.044
7	.050
7.5	.044
8	.030
8.5	.016
9	.007
9.5	.004
10	.011
10.5	.042
11	.129
11.5	.308
12	.576
12.5	.838
13	.950
13.5	.838
14	.576
14.5	.308
15	.129
15.5	.042
16	.011
16.5	.002
17	.000

Mean = 12.700

Standard Deviation = 1.044

Density = .05

Separation = 10

x	f(x)
2	.001
2.5	.002
3	.007
3.5	.016
4	.030
4.5	.044
5	.050
5.5	.044
6	.030
6.5	.016
7	.007
7.5	.002
8	.001
8.5	.000
9	.000
9.5	.000
10	.000
10.5	.000
11	.000
11.5	.002
12	.011
12.5	.042
13	.129
13.5	.308
14	.576
14.5	.838
15	.950
15.5	.838
16	.576
16.5	.308
17	.129
17.5	.042
18	.011
18.5	.002
19	.000

Mean = 14.500

Standard Deviation = 1.118

Density = .10

Separation = 1

x	f(x)
6	.000
6.5	.001
7	.006
7.5	.023
8	.072
8.5	.182
9	.380
9.5	.646
10	.882
10.5	.961
11	.827
11.5	.559
12	.297
12.5	.123
13	.040
13.5	.010
14	.002
14.5	.000

Mean = 10.400

Standard Deviation = 1.005

Density = .10

Separation = 2

x	f(x)
6	.000
6.5	.001
7	.004
7.5	.014
8	.071
8.5	.128
9	.222
9.5	.380
10	.607
10.5	.828
11	.914
11.5	.799
12	.547
12.5	.292
13	.122
13.5	.040
14	.009
14.5	.002
15	.000

Mean = 10.800

Standard Deviation = 1.020

Density = .10

Separation = 4

x	f(x)
5	.000
5.5	.004
6	.014
6.5	.032
7	.061
7.5	.088
8	.100
8.5	.090
9	.071
9.5	.072
10	.135
10.5	.297
11	.547
11.5	.794
12	.900
12.5	.794
13	.546
13.5	.292
14	.122
14.5	.040
15	.000

Mean = 11.600

Standard Deviation = 1.077

Density - .10

Separation = 6

x	f(x)
3.5	.000
4	.001
4.5	.004
5	.014
5.5	.032
6	.061
6.5	.088
7	.100
7.5	.088
8	.061
8.5	.033
9	.014
9.5	.006
10	.011
10.5	.040
11	.123
11.5	.292
12	.546
12.5	.794
13	.900
13.5	.794
14	.546
14.5	.292
15	.122
15.5	.040
16	.009
16.5	.002
17	.000

Mean = 12.400

Standard Deviation = 1.414

Density = .10

Separation = 10

x	f(x)
1.5	.000
2	.001
2.5	.004
3	.014
3.5	.032
4	.061
4.5	.088
5	.100
5.5	.088
6	.061
6.5	.032
7	.014
7.5	.004
8	.001
8.5	.000
9	.000
10	.000
10.5	.000
11	.000
11.5	.002
12	.010
12.5	.040
13	.121
13.5	.292
14	.546
14.5	.794
15	.900
15.5	.794
16	.546
16.5	.292
17	.121
17.5	.040
18	.009
18.5	.002
19	.000

Mean = 14.000

Standard Deviation = 1.414

Density = .20

Separation = 1

x	f(x)
6	.000
6.5	.002
7	.011
7.5	.036
8	.100
8.5	.230
9	.436
9.5	.685
10	.882
10.5	.921
11	.771
11.5	.512
12	.269
12.5	.110
13	.036
13.5	.009
14	.002
14.5	.000

Mean = 10.300

Standard Deviation = 1.020

Density = .20

Separation = 2

x	f(x)
5.5	.000
6	.002
6.5	.009
7	.027
7.5	.067
8	.130
8.5	.212
9	.308
9.5	.436
10	.607
10.5	.771
11	.827
11.5	.715
12	.487
12.5	.260
13	.108
13.5	.035
14	.009
14.5	.002
15	.000

Mean = 10.600

Standard Deviation = 1.078

Density = .20

Separation = 4

x	f(x)
4.5	.000
5	.002
5.5	.009
6	.027
6.5	.064
7	.121
7.5	.177
8	.200
8.5	.178
9	.130
9.5	.100
10	.135
10.5	.269
11	.487
11.5	.706
12	.800
12.5	.706
13	.485
13.5	,260
14	.108
14.5	.035
15	.009
15.5	.002
16	.000

Mean = 11.200

Standard Deviation = 1.281

Density = .20

Separation = 6

x	f(x)
4	.002
4.5	.009
5	.027
5.5	.064
6	.121
6.5	.176
7	.200
7.5	.176
8	.121
8.5	.065
9	.027
9.5	.011
10	.011
10.5	.036
11	.108
11.5	.260
12	.485
12.5	.706
13	.800
13.5	.706
14	.485
14.5	.260
15	.108
15.5	.035
16	.009
16.5	.002
17	.000

Mean = 11.800

Standard Deviation = 1.562

Density = .20

Separation = 10

x	f(x)
1.5	.000
2	.002
2.5	.009
3	.027
3.5	.064
4	.121
4.5	.176
5	.200
5.5	.176
6	.121
6.5	.064
7	.027
7.5	.009
8	.002
8.5	.000
9	.000
9.5	.000
10	.000
10.5	.000
11	.000
11.5	.002
12	.009
12.5	.035
13	.108
13.5	.260
14	.485
14.5	.706
15	.800
15.5	.706
16	.485
16.5	.260
17	.108
17.5	.035
18	.009
18.5	.002
19	.000

Mean = 13.000

Standard Deviation = 2.236

Density = .40

Separation = 1

x	f(x)
5.5	.000
6	.001
6.5	.005
7	.019
7.5	.061
8	.156
8.5	.324
9	.545
9.5	.764
10	.882
10.5	.842
11	.659
11.5	.418
12	.212
12.5	.086
13	.027
13.5	.007
14	.001
14.5	.000

Mean = 10.100

Standard Deviation = 1.078

Density = .40

Separation = 2

x	f(x)
5	.000
5.5	.001
6	.004
6.5	.018
7	.054
7.5	.131
8	.249
8.5	.379
9	.491
9.5	.548
10	.606
10.5	.659
11	.654
11.5	.547
12	.368
12.5	.196
13	.081
13.5	.026
14	.007
14.5	.001
15	.000

Mean = 10.200

Standard Deviation = 1.281

Density = .40

Separation = 4

x	f(x)
4	.000
4.5	.001
5	.004
5.5	.018
6	.054
6.5	.130
7	.243
7.5	.353
8	.400
8.5	.354
9	.249
9.5	.156
10	.135
10.5	.212
11	.368
11.5	.530
12	.600
12.5	.530
13	.364
13.5	.195
14	.081
14.5	.026
15	.007
15.5	.006
16	.001
16.5	.000

Mean = 10.400

Standard Deviation = 1.887

Density = .40

Separation = 6

x	f(x)
3	.000
3.5	.001
4	.003
4.5	.018
5	.054
5.5	.130
6	.243
6.5	.353
7	.400
7.5	.353
8	.243
8.5	.130
9	.054
9.5	.019
10	.011
10.5	.027
11	.081
11.5	.195
12	.364
12.5	.529
13	.600
13.5	.529
14	.364
14.5	.195
15	.081
15.5	.026
16	.007
16.5	.001
17	.000

Mean = 10.600

Standard Deviation = 2.600

Density = .40

Separation = 10

x	f(x)
1	.000
1.5	.001
2	.008
2.5	.018
3	.054
3.5	.130
4	.243
4.5	.353
5	.400
5.5	.353
6	.242
6.5	.130
7	.054
7.5	.017
8	.004
8.5	.001
9	.000
9.5	.000
10	.000
10.5	.000
11	.000
11.5	.001
12	.007
12.5	.026
13	.081
13.5	.195
14	.364
14.5	.529
15	.600
15.5	.529
16	.364
16.5	.195
17	.081
17.5	.026
18	.007
18.5	.001
19	.000

Mean = 11.000

Standard Deviation = 4.123

Density = .50

Separation = 1

x	f(x)
5.5	.000
6	.001
6.5	.006
7	.023
7.5	.073
8	.184
8.5	.371
9	.604
9.5	.803
10	.882
10.5	.803
11	.604
11.5	.371
12	.184
12.5	.073
13	.023
13.5	.006
14	.001
14.5	.000

Mean = 10.000

Standard Deviation = 1.118

Density = .50

Separation = 2

x	f(x)
5	.000
5.5	.001
6	.006
6.5	.022
7	.068
7.5	.163
8	.309
8.5	.463
9	.568
9.5	.604
10	.607
10.5	.604
11	.568
11.5	.463
12	.309
12.5	.163
13	.068
13.5	.022
14	.006
14.5	.001
15	.000

Mean = 10.000

Standard Deviation = 1.414

Density = .50

Separation = 4

x	f(x)
4	.000
4.5	.001
5	.006
5.5	.022
6	.068
6.5	.162
7	.303
7.5	.441
8	.500
8.5	.442
9	.309
9.5	.184
10	.135
10.5	.184
11	.309
11.5	.442
12	.500
12.5	.441
13	.303
13.5	.162
14	.068
14.5	.022
15	.006
15.5	.001
16	.000

Mean = 10.000

Standard Deviation = 2.236

Density = .50

Separation = 6

x	f(x)
3	.000
3.5	.001
4	.006
4.5	.022
5	.068
5.5	.162
6	.303
6.5	.441
7	.500
7.5	.441
8	.303
8.5	.162
9	.068
9.5	.023
10	.011
10.5	.023
11	.068
11.5	.162
12	.303
12.5	.441
13	.500
13.5	.441
14	.303
14.5	.162
15	.068
15.5	.022
16	.006
16.5	.001
17	.000

Mean = 10.000

Standard Deviation = 3.162

Density = .50

Separation = 10

x	f(x)
1	.000
1.5	.001
2	.006
2.5	.022
3	.068
3.5	.162
4	.303
4.5	.441
5	.500
5.5	.441
6	.303
6.5	.162
7	.068
7.5	.022
8	.006
8.5	.001
9	.000
9.5	.000
10	.000
10.5	.000
11	.000
11.5	.001
12	.006
12.5	.022
13	.068
13.5	.162
14	.303
14.5	.441
15	.500
15.5	.441
16	.303
16.5	.162
17	.068
17.5	.022
18	.006
18.5	.001
19	.000

Mean = 10.000

Standard Deviation = 5.099

APPENDIX C

Applesoft Basic Computer Program Used in the Simulation

```
10      REM LINES 11-13 FOR RANDOM START
11      PRINT "INPUT TIME; 3:30 AS 330"
12      INPUT XX
13      YY = ABS ( SIN (XX)) * 100
14      DIM C (18,2),E(18,2)
17      REM NT=TRIAL COUNTER: MW=#OF SUCCESSES FOR U TEST
18      REM TS=#OF SUCCESSES FOR T-TEST
19      PRINT "INPUT DENSITY,SEPARATION"
20      PRINT "SHIFT, # OF TRIALS"
21      INPUT DN,SP,DI,L1
22      K = DN
23      S = SP
24      NT = 1
25      MW = 0
26      TS = 0
85      PRINT "THIS IS TRIAL NUMBER";NT
95      CC = 1
96      EC = 1
100     C(CC,1) = RND (YY) * 20
110     C(CC,2) = RND (YY)
115     Z = C(CC,1)
116     X6 = ((Z - 10 + (S / 2)) ^ 2 / - 2
117     X7 = ((Z - 10 - (S / 2)) ^ 2 / - 2
118     Z1 - (K * EXP (X6)) + ((1 - K) * EXP (X7))
120     IF Z1 < C(CC,2) THEN 100
130     IF CC = 18 THEN 160
140     CC = CC + 1
150     GOTO 100
160     E(EC,1) = RND (YY) * 20
170     E(EC,2) = RND (YY)
175     Z2 = E(EC,1)
177     X8 = ((Z2 - 10 + (S / 2)) ^ 2) / - 2
178     X9 = ((Z2 - 10 - (S / 2)) ^ 2) / - 2
179     Z3 = (K * EXP (X8)) + ((1 - K) * EXP (X9))
180     IF Z3 < E(EC,2) THEN 160
190     IF EC = 18 THEN 220
195     E(EC,1) = E(EC,1) + DI
200     EC = EC + 1
210     GOTO 160
220     REM STUDENT T
230     REM MEAN VALUES
240     SC = 0
241     SE = 0
250     FOR L = 1 TO 18
260     SC = C(L,1) + SC
270     SE = E(L,1) + SE
280     NEXT L
290     MC = SC / 18
300     ME = SE / 18
302     PRINT "ME= ":ME
304     PRINT "MC= ":MC
310     IF ME < MC THEN 330
320     GOTO 340
```

```
330     PRINT "MEAN EX<MEAN C - AUTOMATIC FAIL"
332     IF NT = L1 THEN 810
333     NT = NT + 1
335     TOGO 85
340     S1 = 0
341     S2 = 0
350     FOR N = 1 TO 18
360     S1 = (E(N,1) - ME)^ 2 + S1
370     S2 = (C(N,1) - MC)^ 2 + S2
380     NEXT N
390     S3 = SQR (S1 / 17)
400     S4 = SQR (S2 / 17)
410     S5 = S3 / SQR (18)
420     S6 = S4 / SQR (18)
430     S7 = SQR (S5^ 2 + S6 ^ 2)
435     PRINT "S7= ":S7
440     T = (ME - MC) / S7
450     IF T> = 1.6918 THEN 480
451     REM CRITICAL VAL;UE IS INEXACT
460     PRINT "STUDENT T FAILS"
470     GOTO 490
480     PRINT "STUDENT T PASSES"
485     TS = TS + 1
490     PRINT "T= ";T
500     REM MANN-WHITNEY U
510     REM SORT
520     FOR A = 1 TO 17
530     B = A + 1
540     FOR D = B TO 18
550     IF C(A,1)< = C(D,1) THEN 590
560     DUMMY = C(D,1)
570     C(D,1) = C(A,1)
580     C(A,1) = DUMMY
590     IF E(A,1)< = E(D,1) THEN 630
600     DUMMY = E(D,1)
610     E(D,1) = E(A,1)
620     E(A,1) = DUMMY
630     NEXT D
640     NEXT A
650     REM CALCULATE U
660     U = 0
670     CC = 1
671     EC = 1
680     IF E(EC,1)< C(CC,1) THEN 710
682     IF CC < 18 THEN 690
684     U = (19 - EC) * 18 + U
686     FOR Z1 = EC TO 18
688     NEXT Z1
689     GOTO 740
690     CC = CC + 1
700     GOTO 680
710     U = U + (CC - 1)
712     IF EC < 18 THEN 720
```

```
714     FOR Z2 = CC TO 18
716     NEXT Z2
717     GOTO 740
720     EC = EC + 1
730     GOTO 680
740     IF U > = 215 THEN 770
750     PRINT "MANN WHITNEY U FAILS, U= ";U
760     GOTO 785
770     PRINT "MANN WHITNEY U PASSES, U=";U
780     MW = MW + 1
785     IF NT = L1 THEN 810
790     NT = NT + 1
800     GOTO 85
810     PRINT "OUT OF ";NT;" TRIALS,"
811     PRINT "STUDENT T PASSED ";TS;" TIMES"
812     PRINT "AND MANN WHITNEY U PASSED ";MW;" TIMES."
9999    END
```

APPENDIX D

Test of Apple-II Plus
Pseudorandom Number Generator

The Result For The Generation of
10,000 Pseudorandom Digits

Digit	Observed Frequency	Expected Frequency
0	979	1000
1	1014	1000
2	1028	1000
3	1029	1000
4	975	1000
5	975	1000
6	975	1000
7	1023	1000
8	1024	1000
9	978	1000

Chi Square = 5.783

df = 9

$p > 0.70$

APPENDIX E

Power Estimation and
Relative Power Difference Charts

Power Estimation Normal Curve

Shift	t-test	U-test	Relative Power Difference for U-test
DI = 0.25	.158	.146	-.012
0.50	.402	.368	-.034
0.75	.712	.684	-.028
1.00	.870	.830	-.040
1.25	.968	.952	-.016
1.50	1.000	.998	-.002
1.75	1.000	1.000	.000

Power Estimation

Density = .05
Separation = 1

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.182	.154	-.028
.50	.330	.288	-.042
.75	.664	.632	-.032
1.00	.810	.804	-.006
1.25	.944	.930	-.014
1.50	.982	.980	-.002
1.75	1.000	1.000	.000

Power Estimation

Density = .10
Separation = 1

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.180	.156	-.024
.50	.374	.346	-.028
.75	.604	.588	-.016
1.00	.820	.790	-.030
1.25	.950	.942	-.008
1.50	.992	.986	-.006
1.75	1.000	1.000	.000

Power Estimation

Density = .20
Separation = 1

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.152	.164	-.012
.50	.356	.354	-.002
.75	.620	.600	-.020
1.00	.812	.796	-.016
1.25	.928	.916	-.012
1.50	.994	.990	-.004
1.75	1.000	.996	-.004

Power Estimation

Density = .40
Separation = 1

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.156	.154	-.002
.50	.344	.332	-.012
.75	.604	.564	-.040
1.00	.794	.788	-.006
1.25	.928	.904	-.024
1.50	.972	.956	-.016
1.75	1.000	.996	-.004
2.00	1.000	1.000	.000

Power Estimation

Density = .50
Separation = 1

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.138	.126	-.012
.50	.368	.324	-.044
.75	.586	.554	-.032
1.00	.788	.780	-.008
1.25	.932	.922	-.010
1.50	.980	.970	-.010
1.75	.994	.994	.000
2.00	1.000	1.000	.000

Power Estimation

Density = .05
Separation = 2

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.168	.164	-.004
.50	.366	.336	-.030
.75	.614	.606	-.008
1.00	.830	.792	-.038
1.25	.954	.946	-.008
1.50	.986	.986	.000
1.75	1.000	1.000	.000

Power Estimation

Density = .10
Separation = 2

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.178	.164	-.014
.50	.332	.322	-.010
.75	.570	.546	-.024
1.00	.792	.785	-.006
1.25	.870	.870	.000
1.50	.986	.986	.000
1.75	1.000	1.000	.000

Power Estimation

Density = .20
Separation = 2

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.142	.142	.000
.50	.292	.290	-.002
.75	.478	.488	.010
1.00	.714	.692	-.022
1.25	.854	.838	-.016
1.50	.970	.962	-.008
1.75	.986	.986	.000
2.00	.998	.998	.000
2.25	1.000	1.000	.000

Power Estimation

Density = .40
Separation = 2

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.144	.154	.010
.50	.254	.238	-.016
.75	.610	.570	-.040
1.00	.624	.584	-.040
1.25	.752	.734	-.018
1.50	.888	.868	-.020
1.75	.944	.914	-.030
2.00	.970	.960	-.010
2.25	1.000	1.000	.000

Power Estimation

Density = .50
Separation = 2

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.120	.108	-.012
.50	.240	.220	-.020
.75	.398	.380	-.018
1.00	.628	.566	-.062
1.25	.786	.742	-.044
1.50	.900	.856	-.044
1.75	.986	.972	-.014
2.00	.988	.976	-.012
2.25	1.000	1.000	.000

Power Estimation

Density = .05
Separation = 4

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.158	.158	.000
.50	.280	.304	.024
.75	.614	.902	.288
1.00	.768	.924	.156
1.25	.840	.922	.082
1.50	.924	.960	.036
1.75	.978	1.000	.024
2.00	1.000	1.000	.000

Power Estimation

Density = .10

Separation = 4

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.170	.210	.040
.50	.280	.350	.070
.75	.410	.450	.040
1.00	.540	.630	.090
1.25	.680	.800	.120
1.50	.800	.854	.054
1.75	.958	.972	.014
2.00	.970	.990	.020
2.25	.994	.994	.000
2.50	1.000	1.000	.000

Power Estimation

Density = .20
Separation = 4

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.096	.116	.020
.50	.176	.212	.036
.75	.306	.390	.084
1.00	.422	.558	.136
1.25	.574	.698	.124
1.50	.718	.814	.096
1.75	.824	.886	.062
2.00	.900	.934	.034
2.25	.968	.976	.008
2.50	.980	.984	.004
2.75	1.000	1.000	.000

Power Estimation

Density = .40
Separation = 4

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.118	.122	.004
.50	.116	.124	.008
.75	.232	.268	.036
1.00	.344	.386	.042
1.25	.480	.516	.036
1.50	.470	.510	.040
1.75	.666	.674	.008
2.00	.826	.786	-.040
2.25	.890	.834	-.056
2.50	.934	.886	-.048
2.75	.964	.924	-.040
3.00	.978	.938	-.040
3.25	1.000	.962	-.038

Power Estimation

Density = .50
Separation = 4

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.098	.104	.006
.50	.178	.184	.008
.75	.202	.264	.062
1.00	.316	.346	.030
1.25	.450	.478	.028
1.50	.556	.576	.020
1.75	.720	.694	-.026
2.00	.796	.768	-.028
2.25	.884	.842	-.040
2.50	.928	.868	-.060
2.75	.964	.892	-.072
3.00	.976	.942	-.034
3.25	.986	.972	-.016
3.50	.994	.962	-.032
3.75	.998	.978	-.020
4.00	1.000	1.000	.000

Power Estimation

Density = .05
Separation = 6

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.142	.148	.006
.50	.260	.360	.100
.75	.420	.584	.164
1.00	.568	.768	.200
1.25	.716	.902	.186
1.50	.846	.960	.114
1.75	.916	.986	.070
2.00	.944	.988	.044
2.25	.992	1.000	.008
2.50	1.000	1.000	.000

Power Estimation

Density = .10
Separation = 6

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.120	.142	.022
.50	.200	.250	.050
.75	.220	.413	.193
1.00	.342	.584	.242
1.25	.506	.820	.314
1.50	.666	.930	.264
1.75	.804	.960	.156
2.00	.910	.986	.076
2.25	.956	.994	.038
2.50	.952	.996	.044
2.75	.966	.994	.028
3.00	.984	.998	.014
3.25	.988	1.000	.012

Power Estimation

Density = .20
Separation = 6

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.062	.122	.060
.50	.138	.222	.084
.75	.238	.384	.146
1.00	.308	.504	.196
1.25	.388	.674	.286
1.50	.482	.788	.306
1.75	.588	.848	.260
2.00	.676	.898	.222
2.25	.780	.916	.136
2.50	.826	.932	.106
2.75	.878	.946	.068
3.00	.906	.950	.044
3.25	.952	.972	.020
3.50	.960	.966	.006
3.75	.976	.970	-.006
4.00	.988	.982	-.006
4.25	.996	.990	-.006
4.50	1.000	.990	-.010

Power Estimation

Density = .40
Separation = 6

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.088	.106	.018
.50	.100	.164	.064
.75	.142	.246	.104
1.00	.202	.330	.128
1.25	.268	.478	.210
1.50	.346	.588	.242
1.75	.424	.652	.228
2.00	.538	.736	.198
2.25	.638	.774	.136
2.50	.726	.798	.072
2.75	.772	.816	.042
3.00	.866	.860	-.006
3.25	.924	.862	-.062
3.50	.942	.870	-.072
3.75	.960	.876	-.084
4.00	.982	.910	-.072
4.25	.988	.922	-.066
4.50	.986	.942	-.044
4.75	.994	.948	-.046
5.00	.996	.970	-.026
5.25	1.000	.982	-.018

Power Estimation

Density = .50
Separation - 6

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.070	.096	.026
.50	.100	.144	.044
.75	.152	.216	.064
1.00	.240	.348	.108
1.25	.276	.440	.164
1.50	.376	.552	.176
1.75	.442	.608	.166
2.00	.512	.674	.162
2.25	.618	.738	.120
2.50	.698	.748	.050
2.75	.764	.778	.014
3.00	.810	.788	-.022
3.25	.866	.812	-.044
3.50	.898	.818	-.080
3.75	.940	.830	-.110
4.00	.992	.864	-.128
4.25	.986	.853	-.133
4.50	.990	.896	-.094
4.75	.996	.936	-.060
5.00	1.000	.952	-.048

Power Estimation

Density = .05
Separation = 10

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.101	.152	.051
.50	.190	.292	.102
.75	.398	.584	.186
1.00	.364	.732	.368
1.25	.540	.900	.360
1.50	.628	.972	.344
1.75	.722	.994	.272
2.00	.770	.990	.220
2.25	.860	1.000	.140
2.50	.876	.996	.120
2.75	.892	1.000	.108
3.00	.934	1.000	.064
3.25	.958	.998	.040
3.50	.978	.998	.020

Power Estimation

Density = .10
Separation = 10

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.084	.116	.032
.50	.130	.294	.164
.75	.186	.496	.310
1.00	.250	.644	.394
1.25	.326	.810	.484
1.50	.352	.884	.532
1.75	.526	.944	.418
2.00	.530	.958	.428
2.25	.594	.976	.382
2.50	.696	.990	.294
2.75	.746	.988	.242
3.00	.838	.992	.154
3.25	.870	.992	.122
3.50	.924	1.000	.076

Power Estimation

Density = .20
Separation = 10

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.078	.122	.044
.50	.080	.236	.156
.75	.128	.394	.266
1.00	.160	.522	.362
1.25	.184	.654	.470
1.50	.256	.758	.502
1.75	.302	.848	.546
2.00	.372	.886	.514
2.25	.444	.924	.480
2.50	.494	.926	.432
2.75	.590	.954	.364
3.00	.662	.962	.300
3.25	.690	.940	.250
3.50	.756	.964	.208
3.75	.826	.968	.142
4.00	.855	.980	.125
4.25	.884	.960	.076
4.50	.886	.954	.068
4.75	.924	.964	.040
5.00	.952	.970	.018
5.25	.962	.960	-.002
5.50	.990	.986	-.004
5.75	1.000	.984	-.016

Power Estimation

Density = .40
Separation = 10

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.054	.094	.040
.50	.070	.166	.096
.75	.088	.256	.168
1.00	.126	.356	.230
1.25	.148	.468	.320
1.50	.184	.590	.406
1.75	.242	.652	.410
2.00	.302	.718	.416
2.25	.334	.766	.432
2.50	.386	.790	.404
2.75	.430	.824	.396
3.00	.498	.836	.338
3.25	.550	.850	.300
3.50	.610	.858	.248
3.75	.646	.878	.232
4.00	.692	.850	.158
4.25	.778	.840	.062
4.50	.818	.868	.050
4.75	.856	.858	.002
5.00	.914	.882	-.032
5.25	.920	.868	-.052
5.50	.942	.872	-.070
5.75	.960	.852	-.108

Power Estimation

Density = .50
Separation = 10

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.044	.070	.026
.50	.040	.084	.044
.75	.072	.262	.192
1.00	.110	.350	.240
1.25	.168	.432	.264
1.50	.178	.514	.336
1.75	.208	.572	.364
2.00	.284	.714	.430
2.25	.340	.740	.400
2.50	.376	.764	.388
2.75	.402	.810	.408
3.00	.468	.812	.344
3.25	.508	.796	.288
3.50	.584	.850	.266
3.75	.678	.856	.178
4.00	.696	.832	.136
4.25	.746	.832	.086
4.50	.832	.852	.020
4.75	.844	.834	-.010
5.00	.880	.838	-.042
5.25	.932	.852	-.080
5.50	.940	.840	-.100
5.75	.960	.838	-.124

Power Estimation

Density = .95
Separation = 33
Sub-Population Standard Deviation Ratio = 1/10

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.07	.13	.06
.50	.09	.24	.15
.75	.14	.37	.23
1.00	.14	.51	.37
1.25	.28	.59	.31
1.50	.30	.69	.39
1.75	.33	.69	.36
2.00	.34	.71	.37
2.25	.37	.80	.43
2.50	.37	.84	.47
2.75	.42	.85	.43
3.00	.40	.88	.48
3.25	.41	.89	.48
3.50	.44	.92	.48
3.75	.44	.88	.48
4.00	.46	.95	.49

Power Estimation

Density = .95
Separation = 33
Sub-Population Standard Deviation Ratio = 1/4

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.08	.11	.03
.50	.14	.28	.14
.75	.15	.37	.22
1.00	.16	.46	.30
1.25	.22	.65	.43
1.50	.24	.71	.47
1.75	.29	.76	.47
2.00	.28	.76	.48
2.25	.33	.81	.48
2.50	.33	.89	.56
2.75	.34	.88	.54
3.00	.33	.88	.55
3.25	.33	.89	.56
3.50	.38	.91	.53
3.75	.43	.95	.52
4.00	.38	.94	.56

Power Estimation

Density = .95
Separation = 33
Sub-Population Standard Deviation Ratio = 1/2

Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.02	.13	.09
.50	.03	.25	.22
.75	.07	.37	.30
1.00	.13	.54	.41
1.25	.25	.56	.31
1.50	.19	.67	.48
1.75	.32	.71	.39
2.00	.31	.75	.44
2.25	.31	.87	.46
2.50	.30	.90	.60
2.75	.33	.90	.57
3.00	.32	.89	.57
3.25	.36	.87	.51
3.50	.38	.91	.53
3.75	.37	.92	.54
4.00	.36	.95	.59

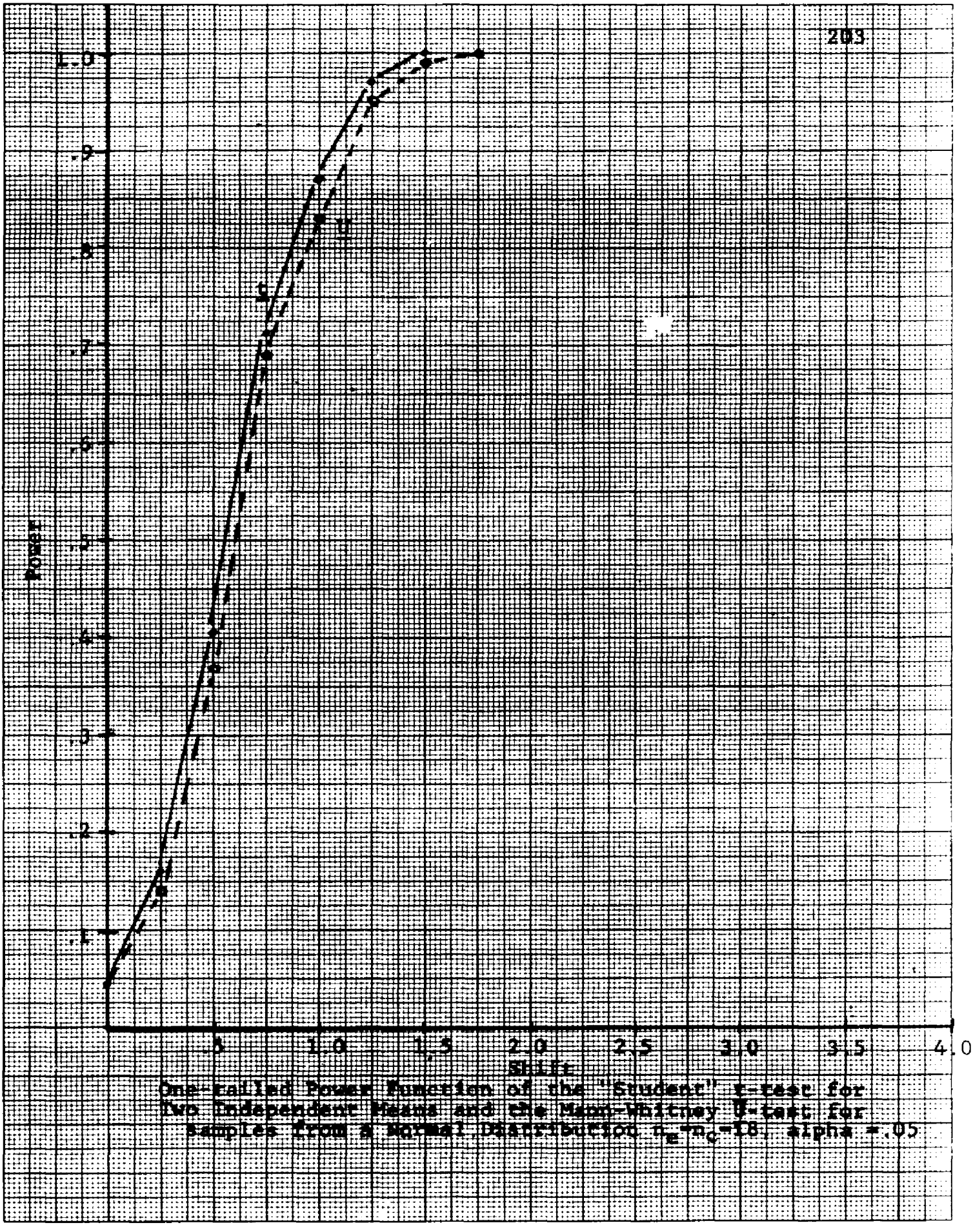
Power Estimation

Density = .95
Separation = 33
Sub-Population Standard Deviation Ratio = 1/1

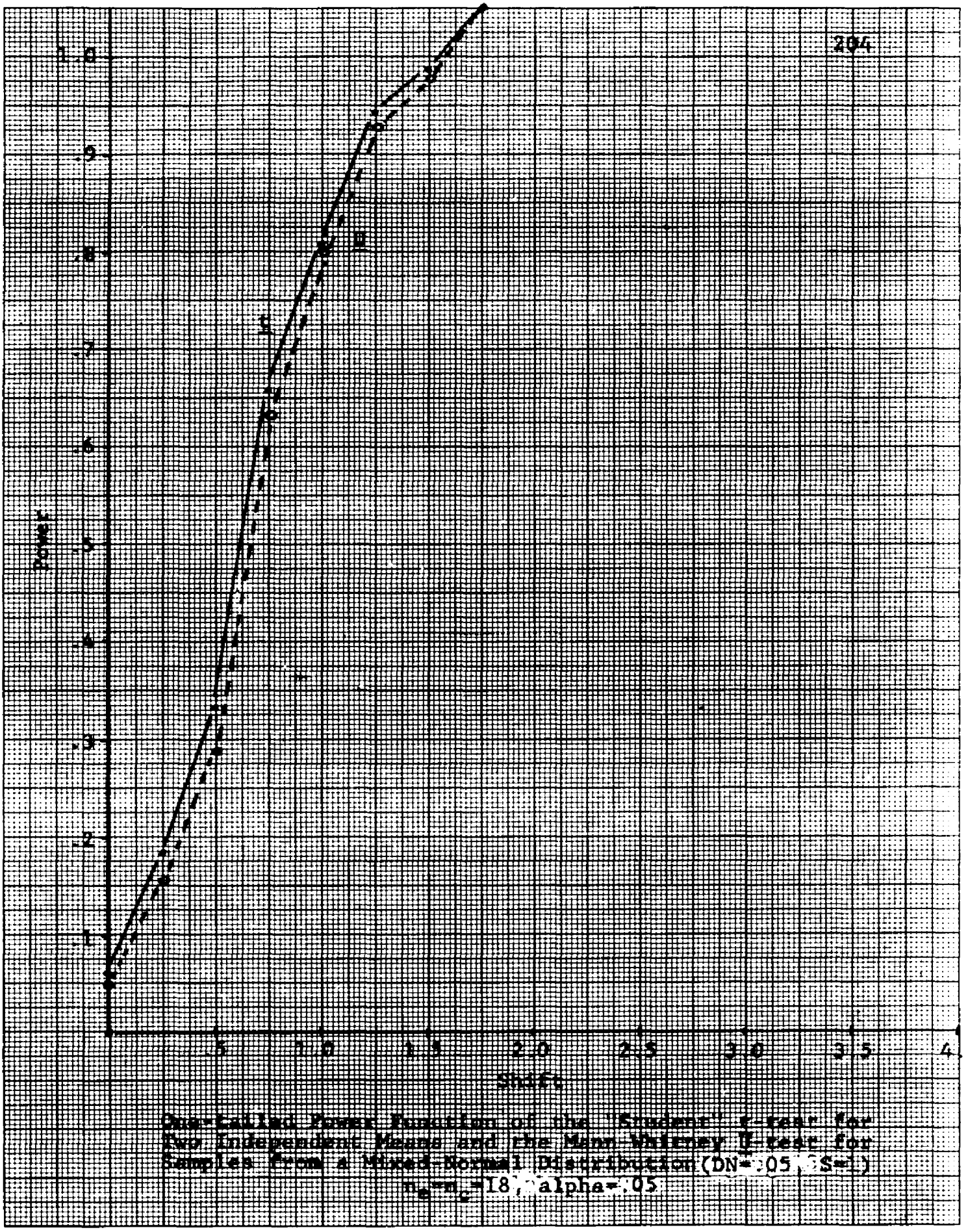
Shift	t-test	U-test	Relative Power Difference for U-test
DI = .25	.02	.25	.23
.50	.05	.29	.24
.75	.11	.43	.22
1.00	.20	.51	.31
1.25	.25	.63	.38
1.50	.30	.62	.32
1.75	.30	.71	.41
2.00	.33	.77	.44
2.25	.36	.78	.42
2.50	.32	.54	.52
2.75	.38	.86	.48
3.00	.38	.84	.46
3.25	.42	.85	.43
3.50	.36	.84	.48
3.75	.40	.92	.52
4.00	.39	.91	.52

APPENDIX F

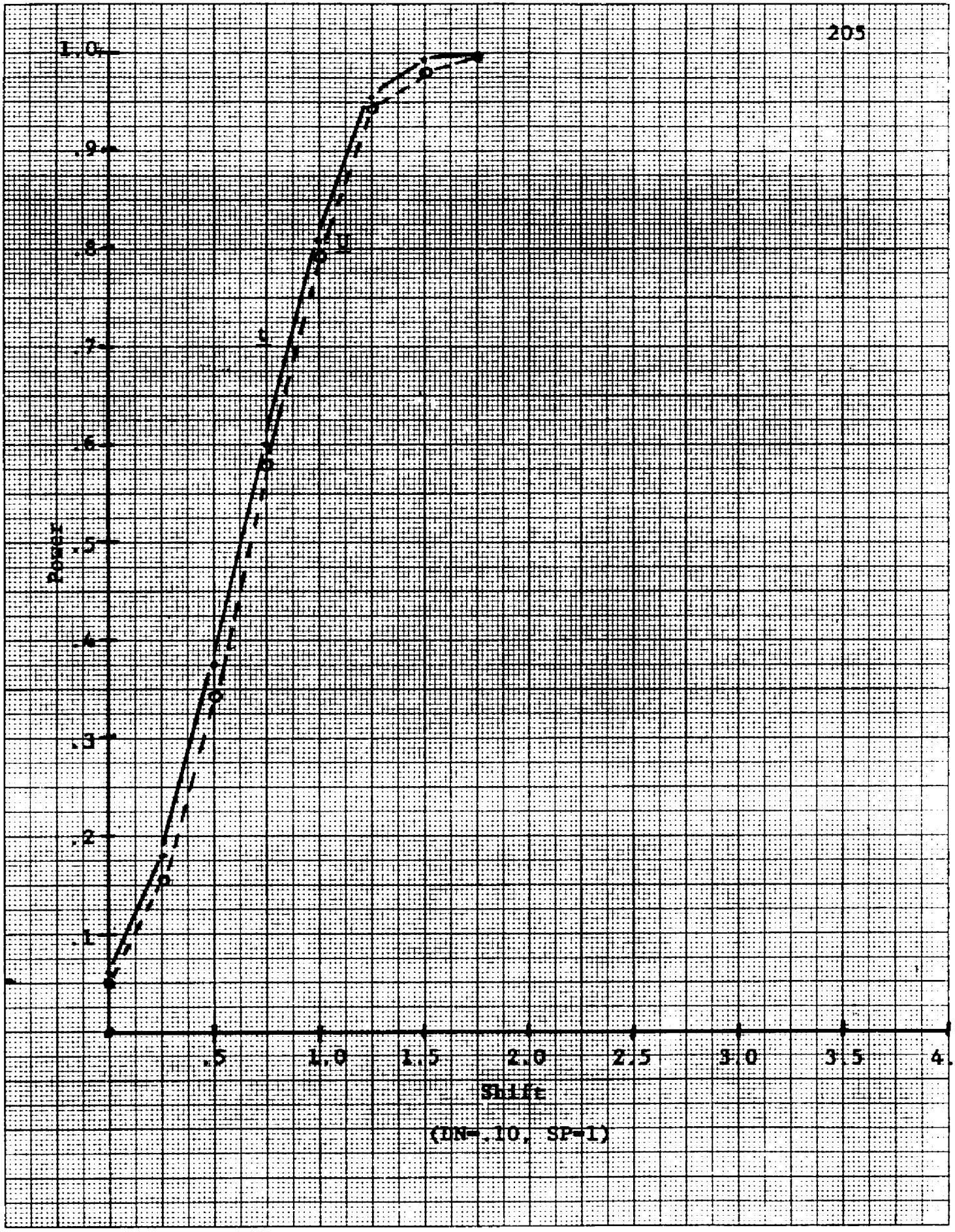
Power Function Graphs

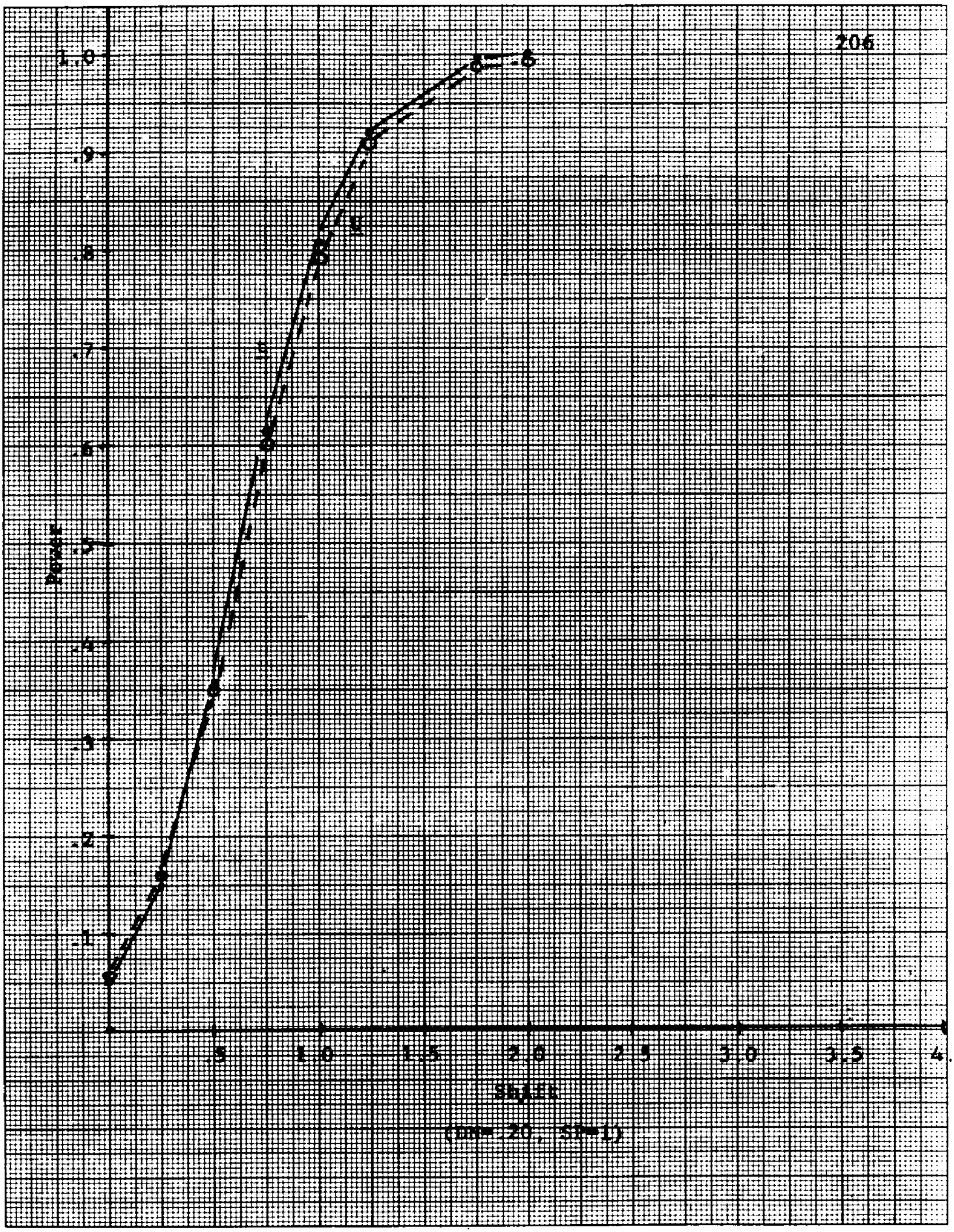


One-tailed Power Function of the "Student" t-test for two Independent Means and the Mann-Whitney U-test for samples from a Normal Distribution $n_e = n_c = 18$, alpha = .05

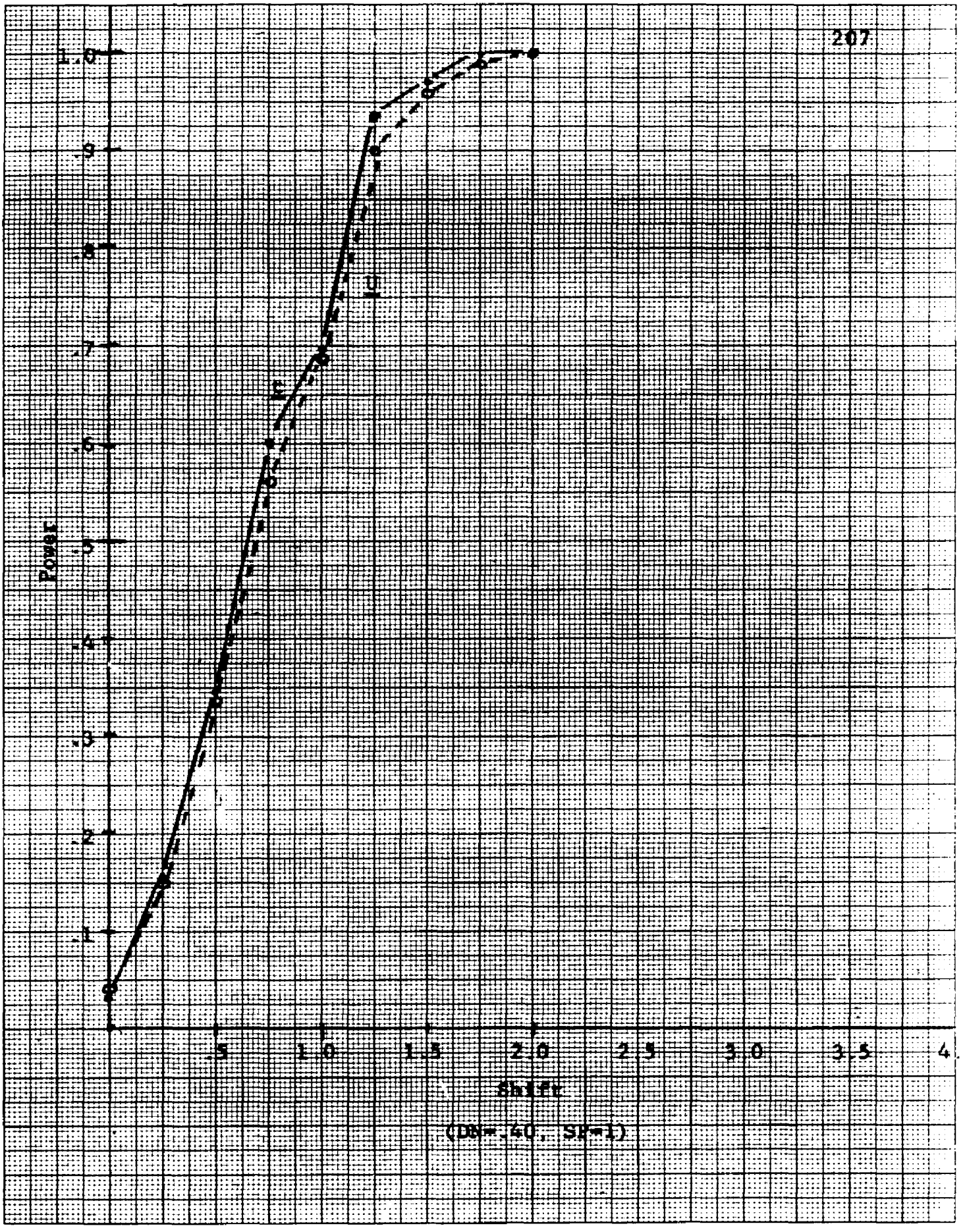


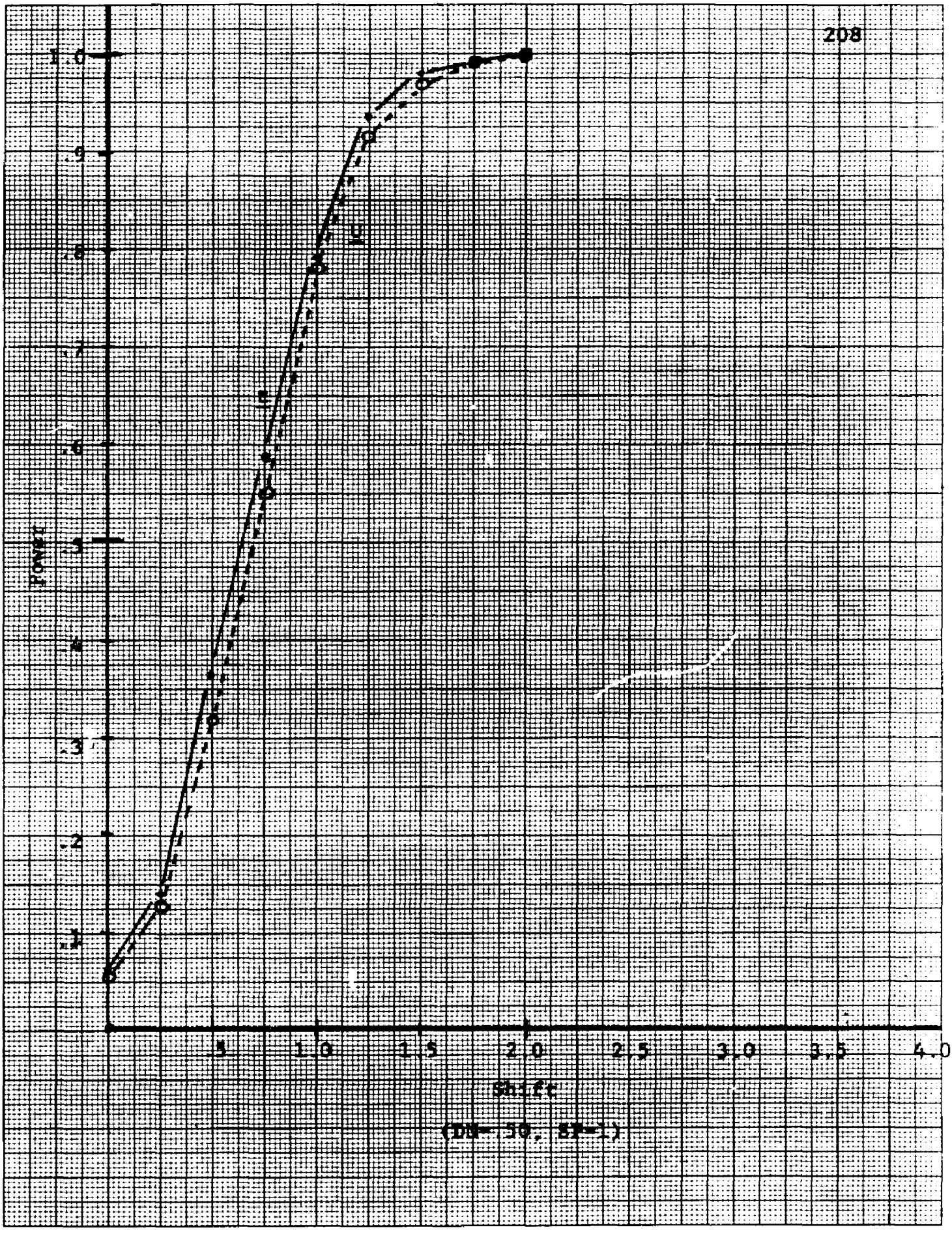
One-Tailed Power Function of the "Student" t-test for Two Independent Means and the Mann-Whitney U-test for Samples from a Mixed-Normal Distribution(DN=.05, S=1)
$n_e = n_c = 18$, alpha=.05


1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
Power
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Shift
(DN=.10, SP=1)

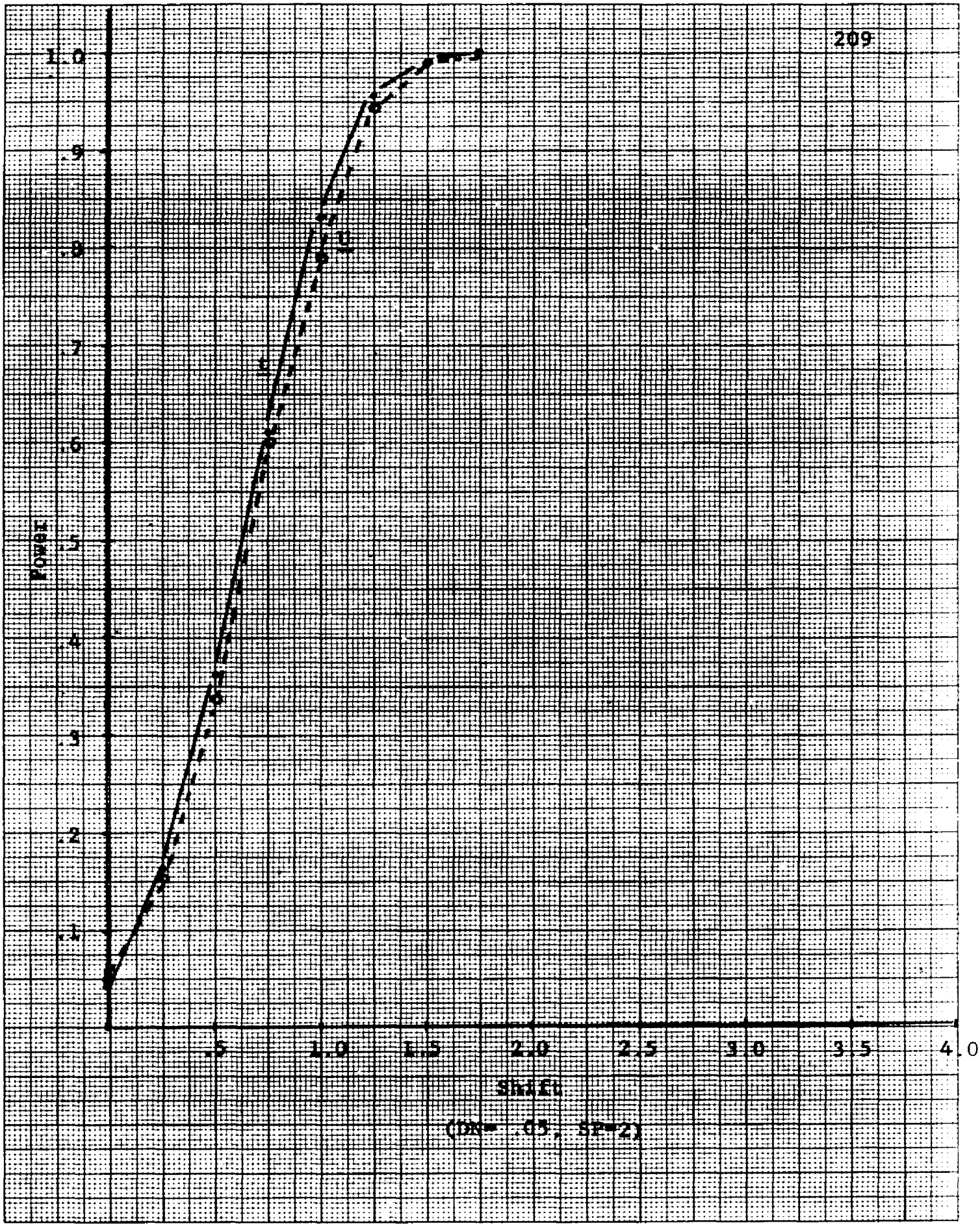


Millimeters to the Centimeter

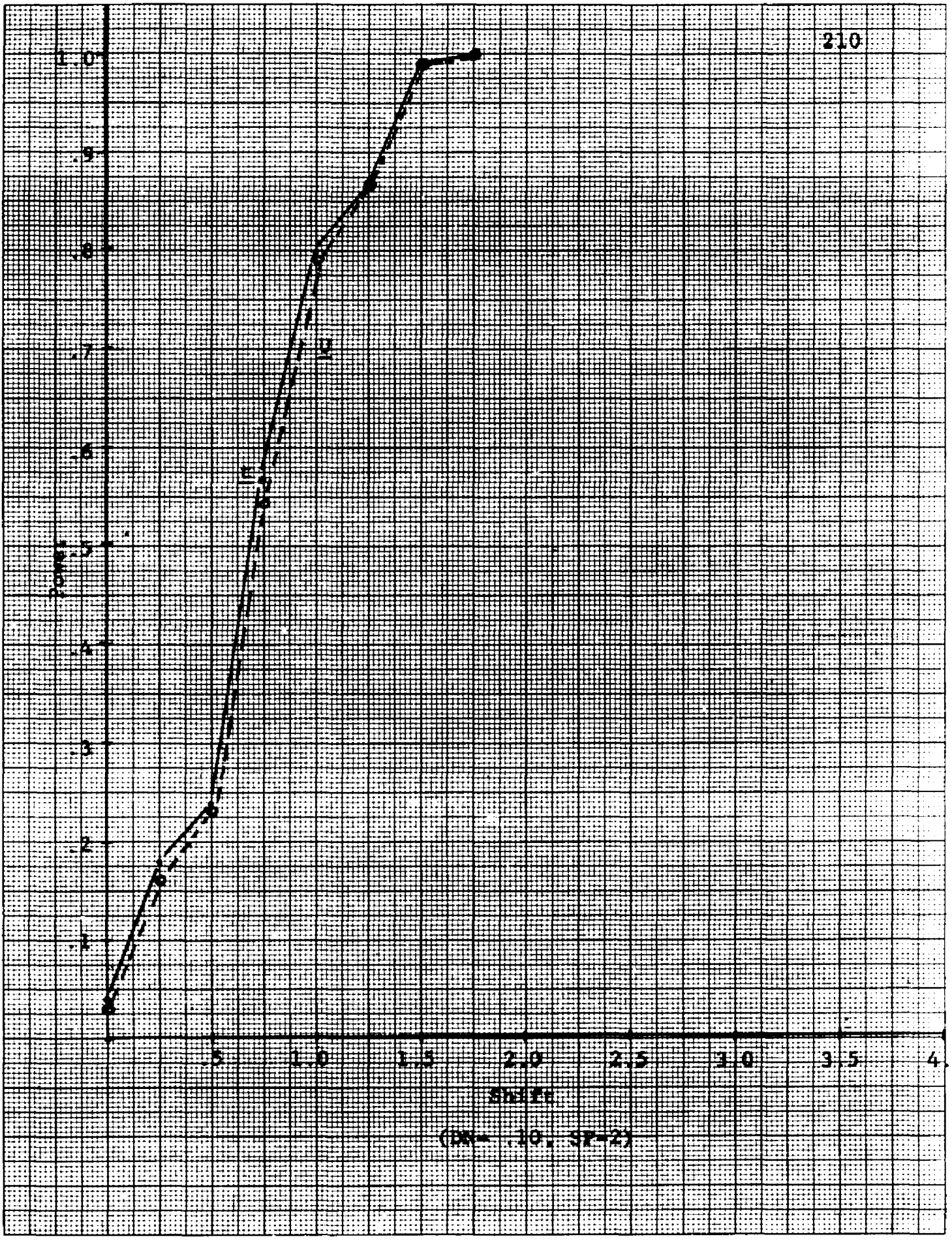

1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
Power
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Shift
(DN=.40, SH=1)



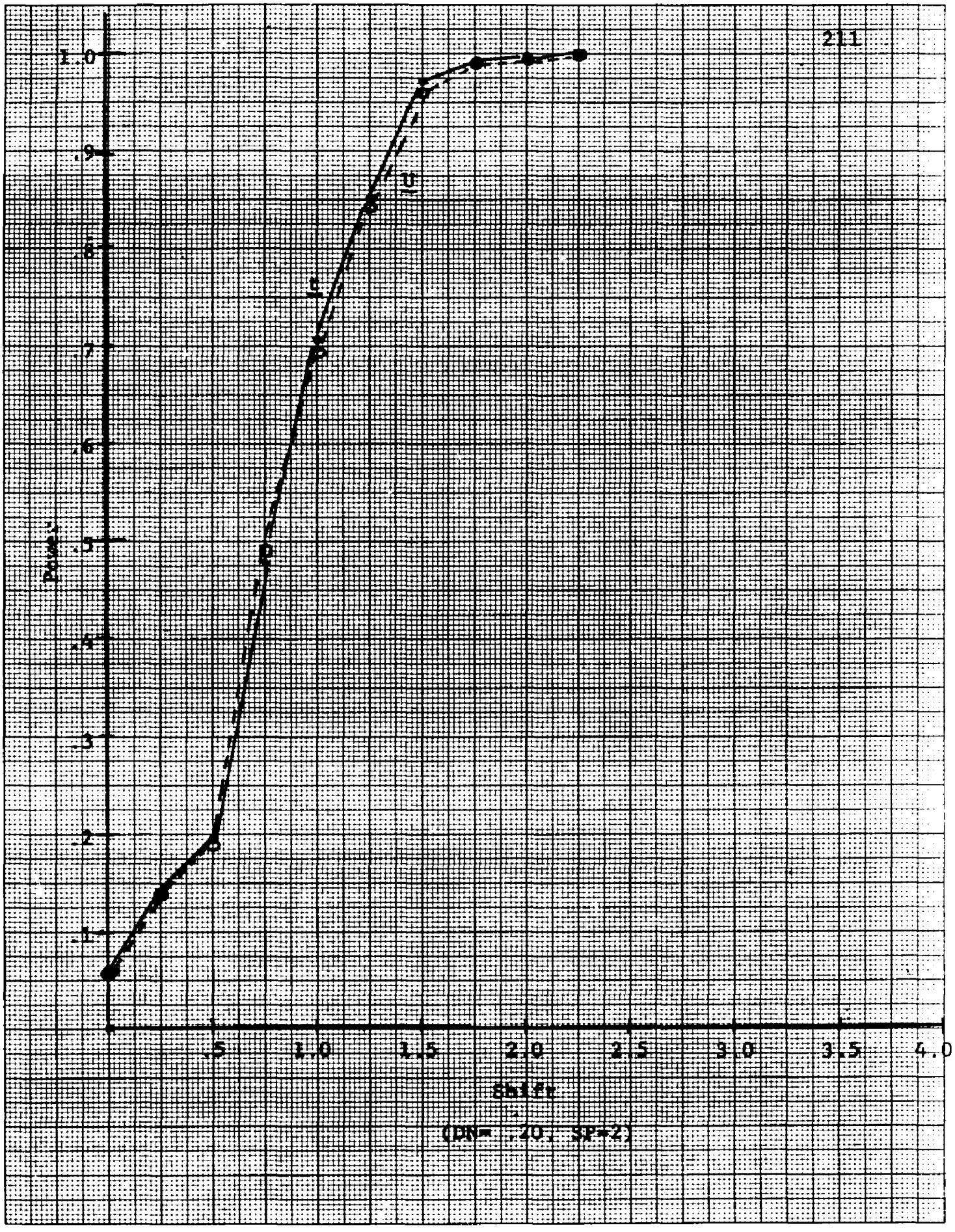
Millimeters to the Centimeter

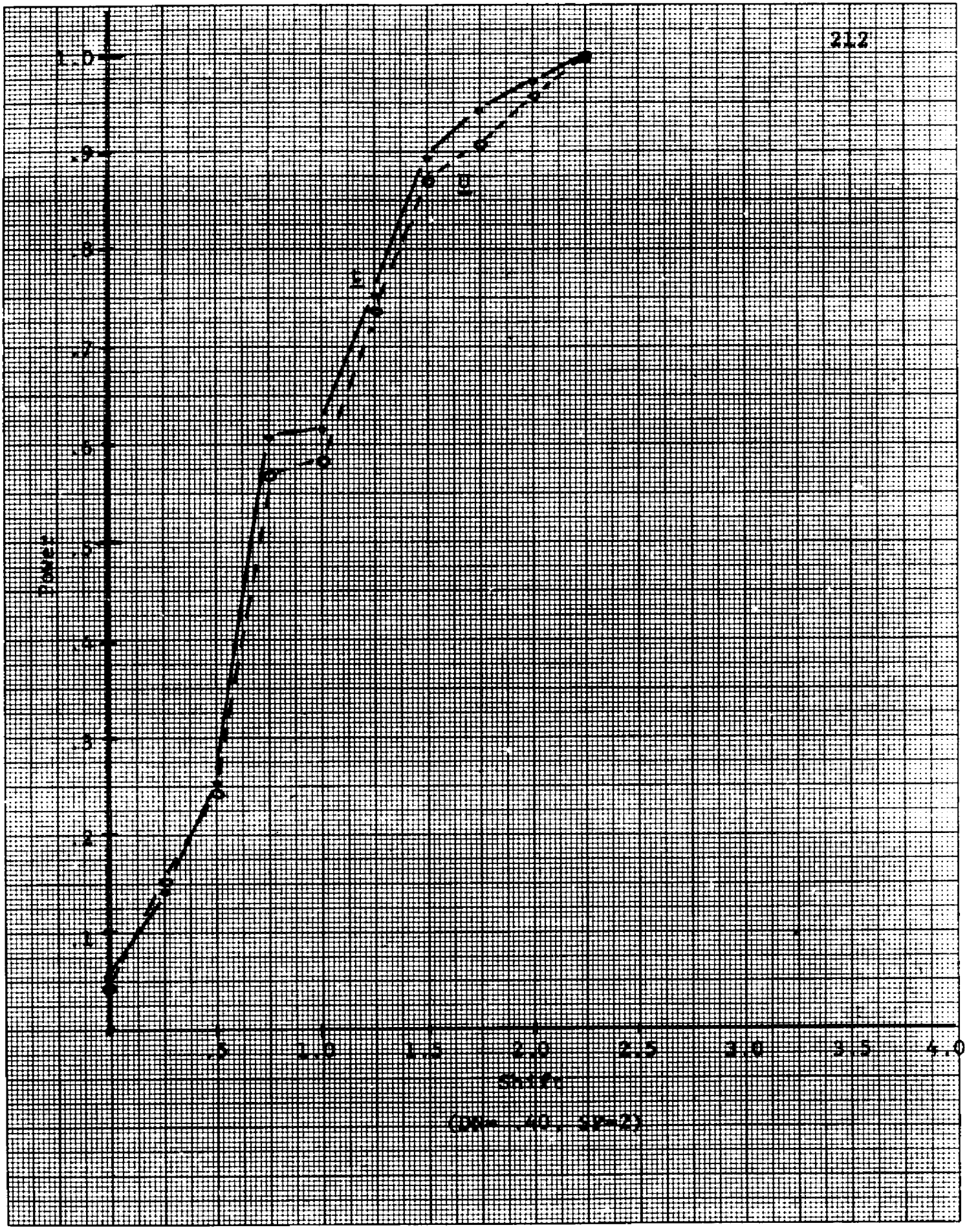


Millimeters to the Centimeter

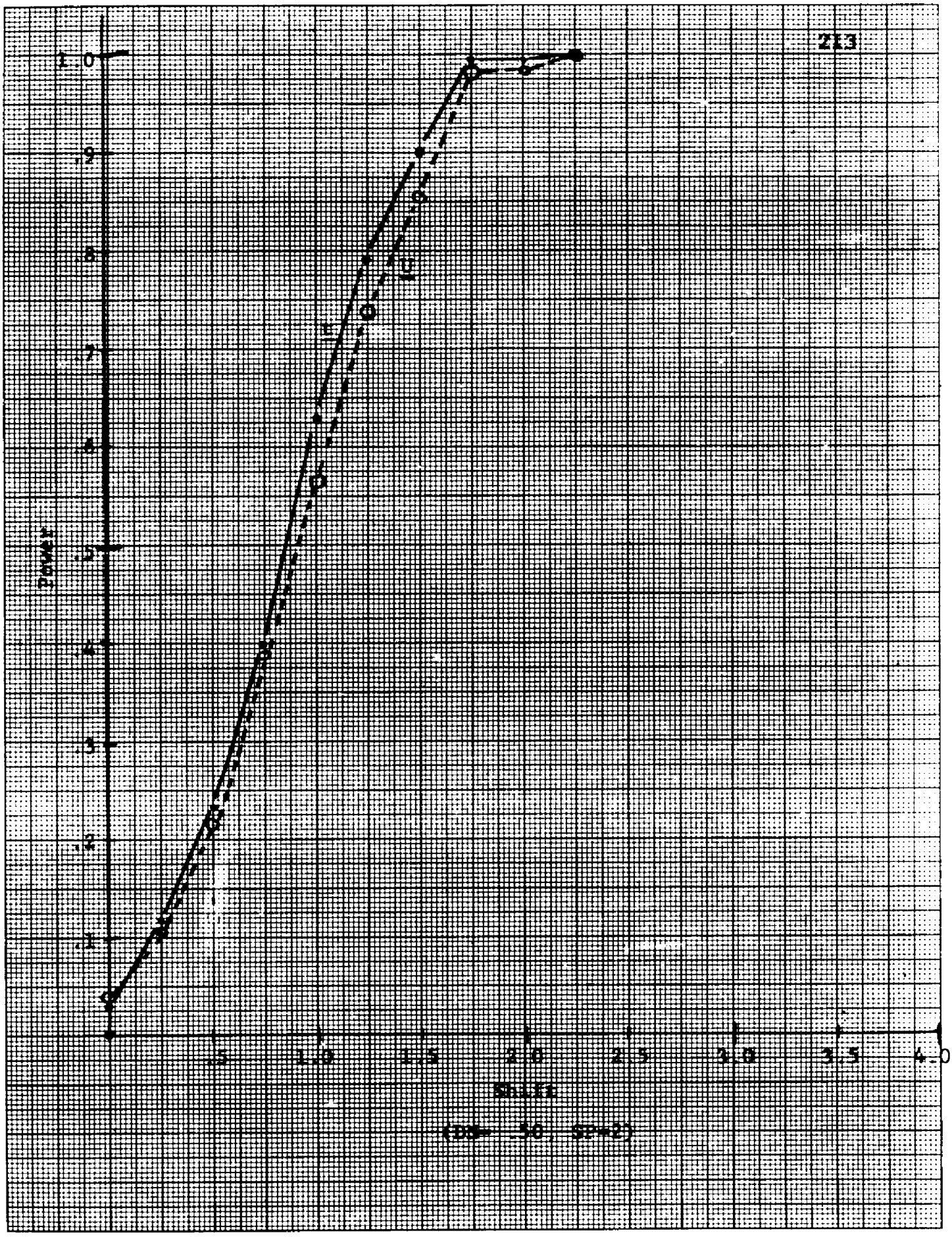


Millimeters to the Centimeter

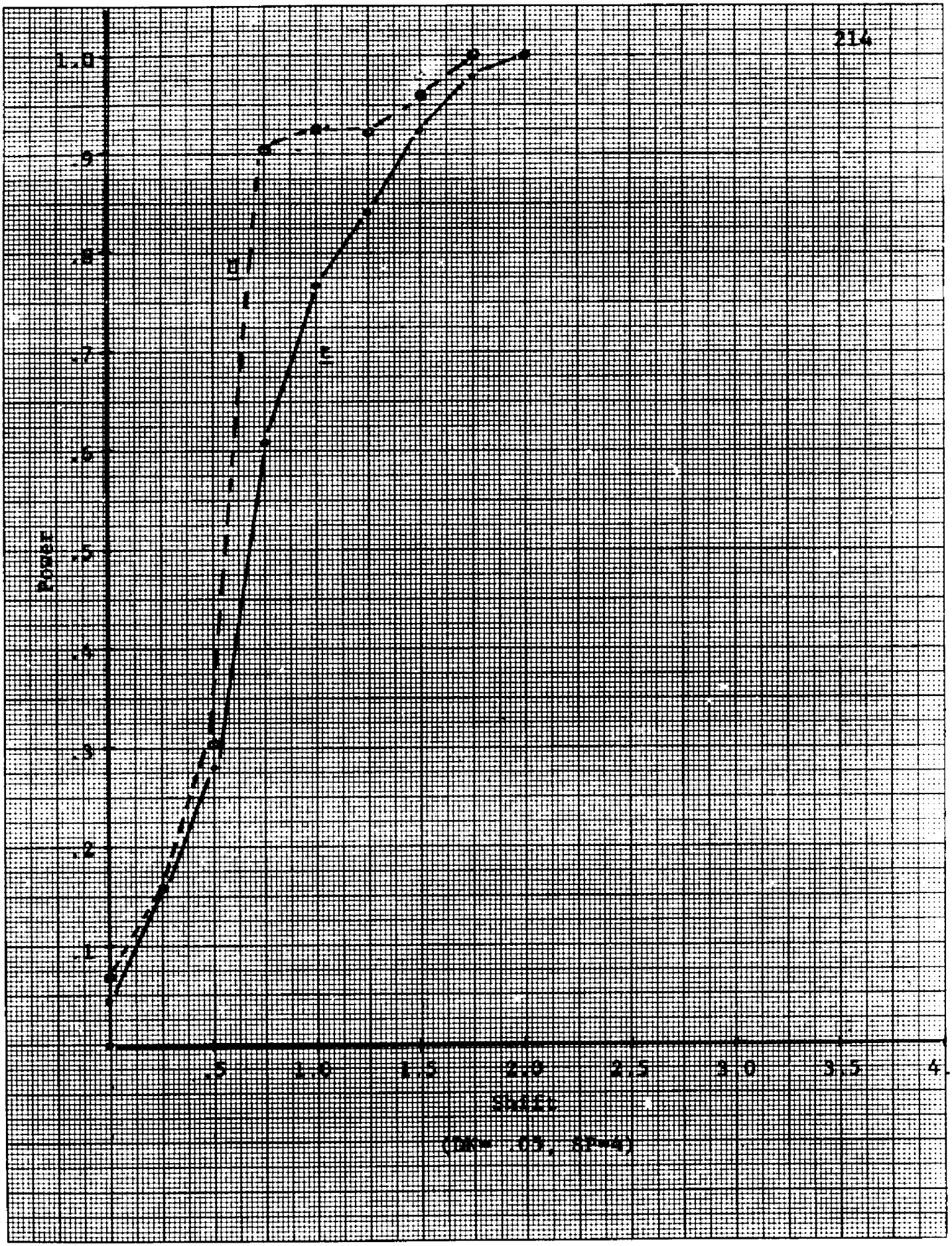

211
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
Ps
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
δh(ft
(DN= .20, SP=2)



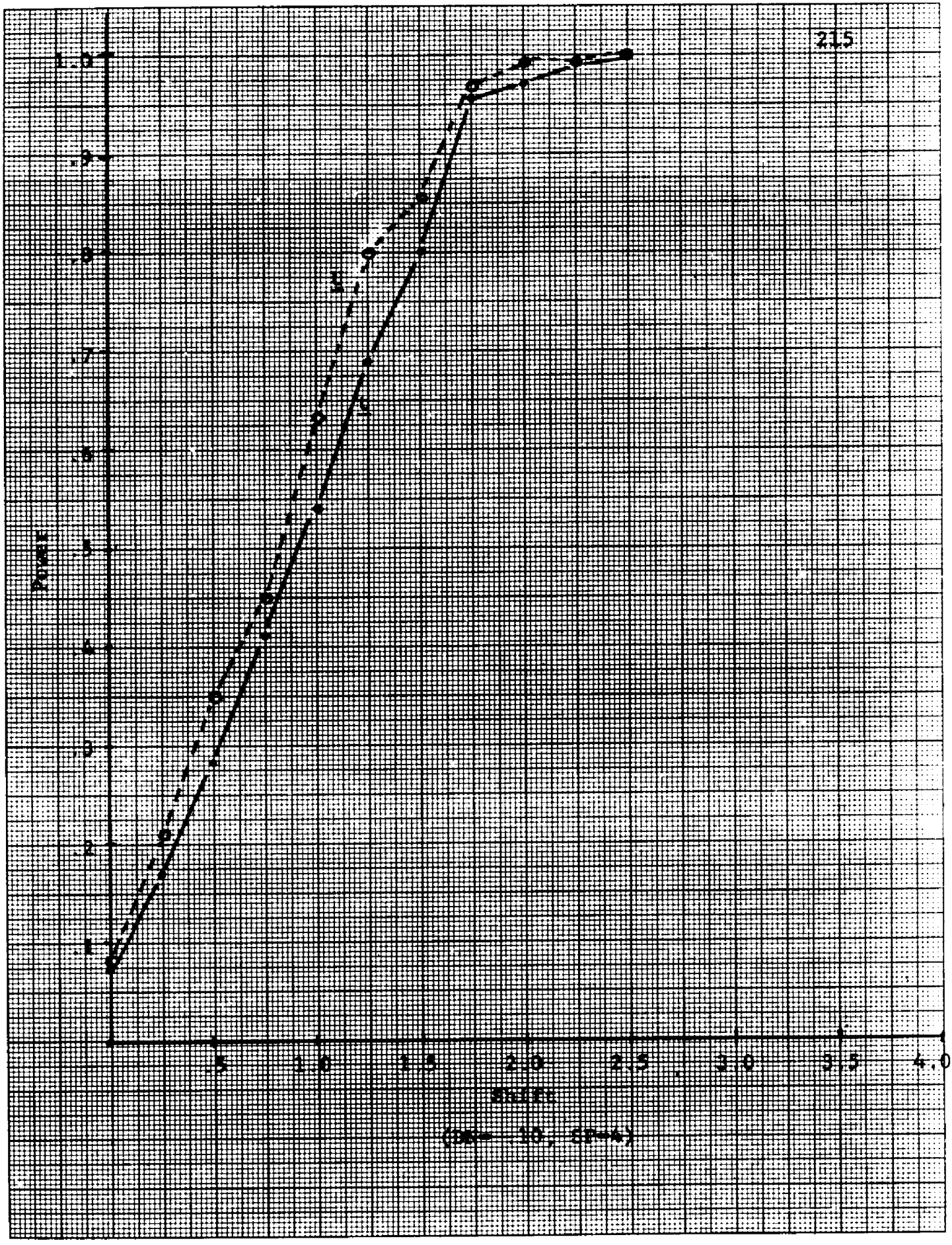
imeters to the Centimeter

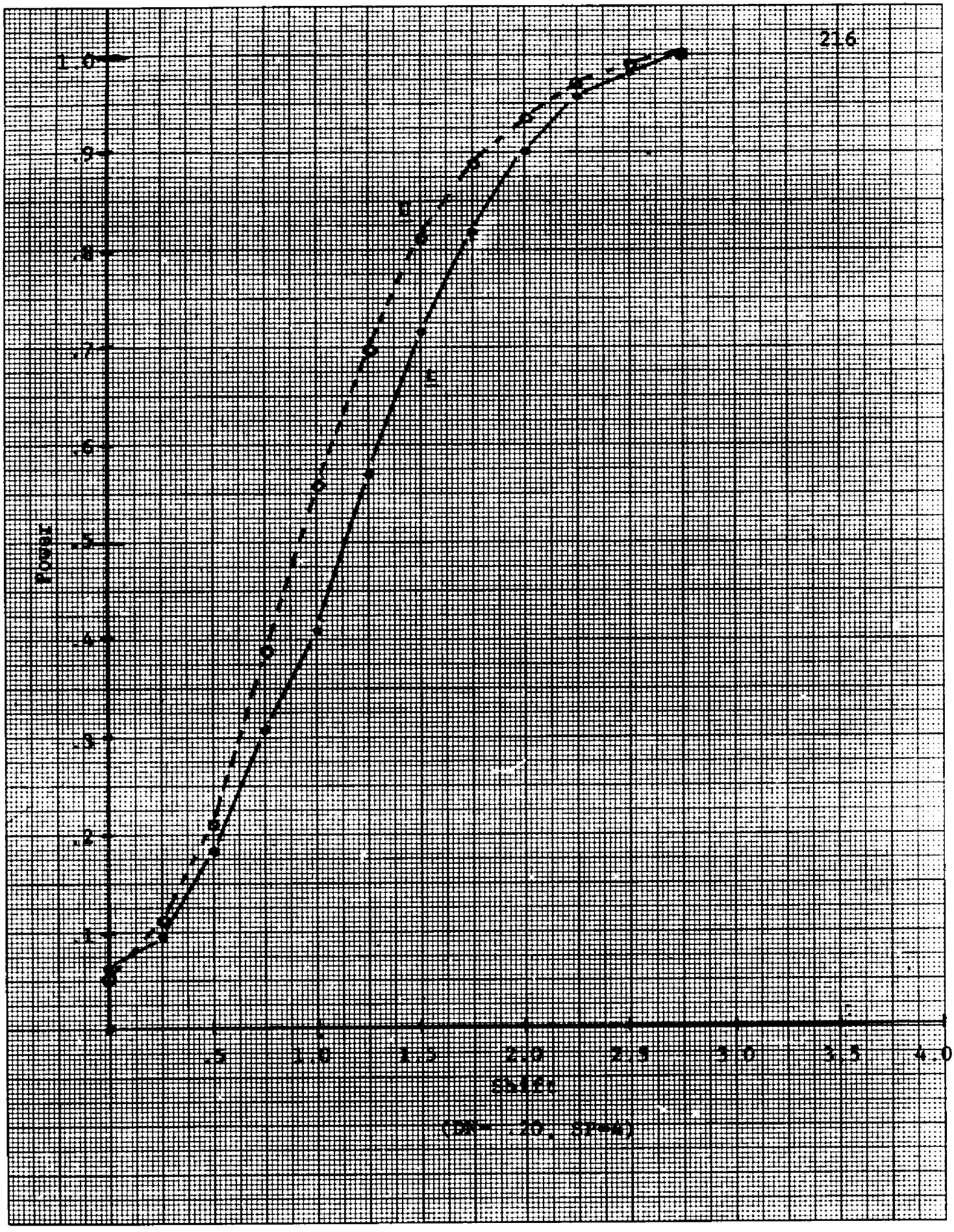


Millimeters to the Centimeter

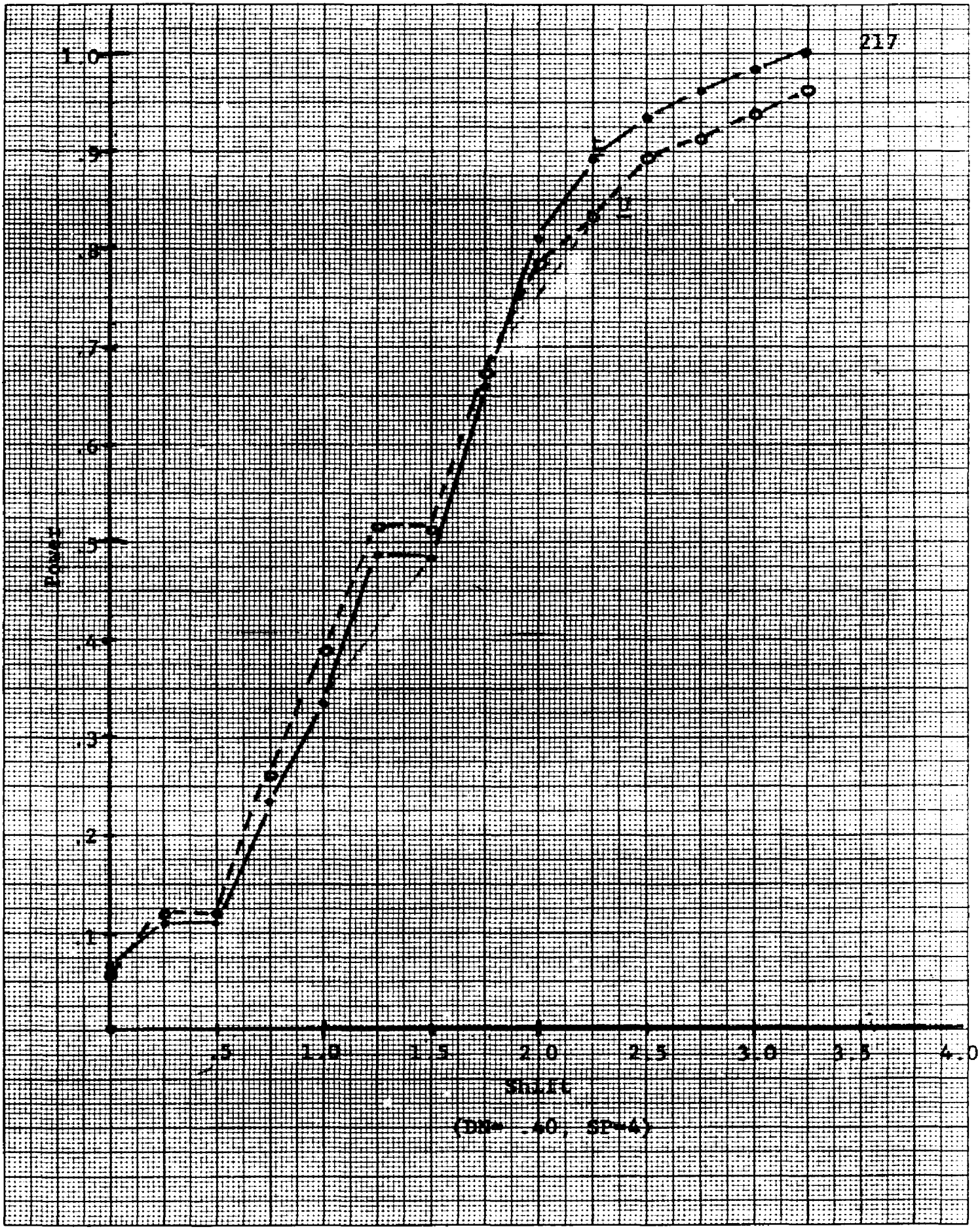


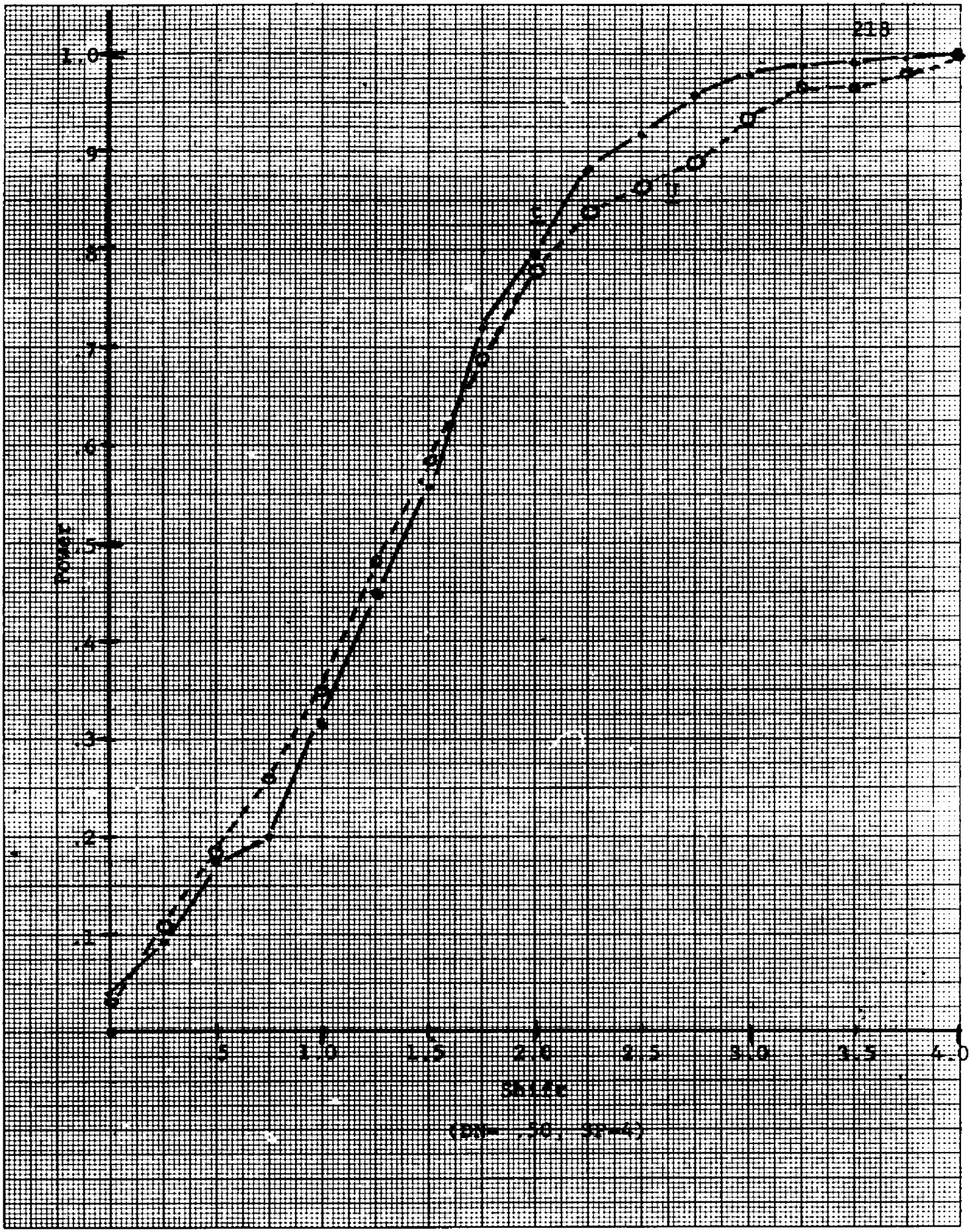
Millimeters to the Centimeter


21.5
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Shift
(DB=.10, SF=4)

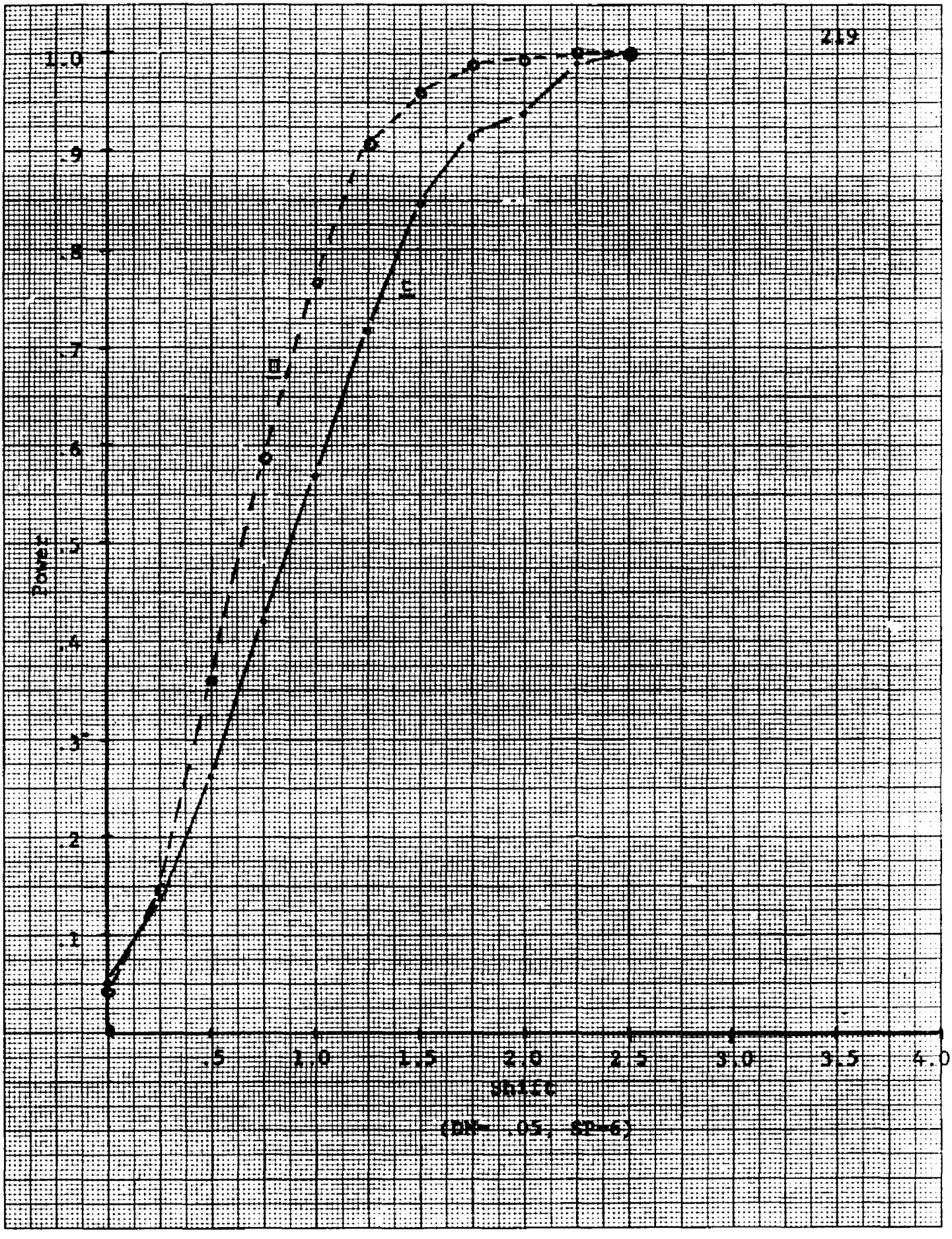


Millimeters to the Centimeter

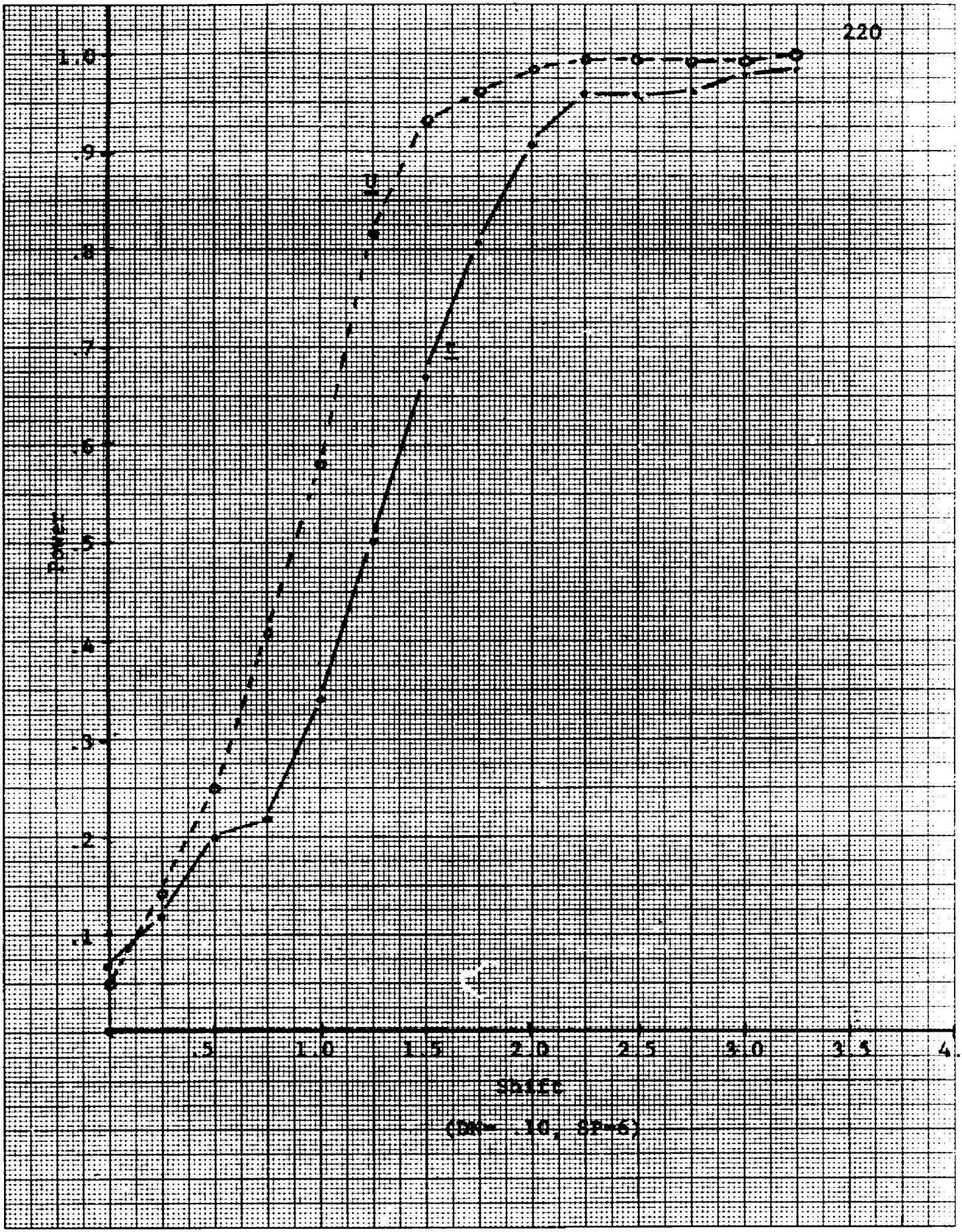

1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
Ratio
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Shift
(DS= .40, SP=4)



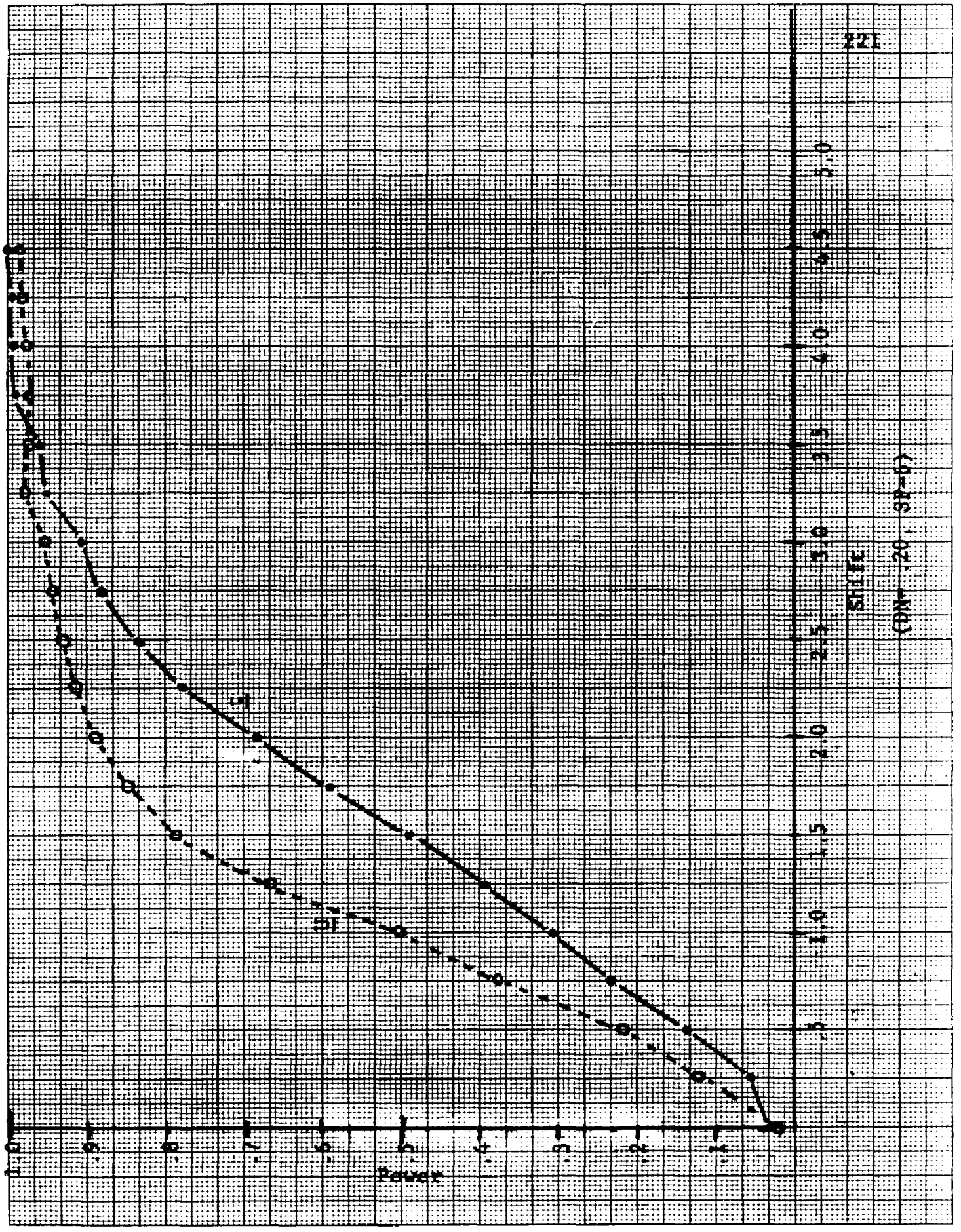
Millimeters to the Centimeter



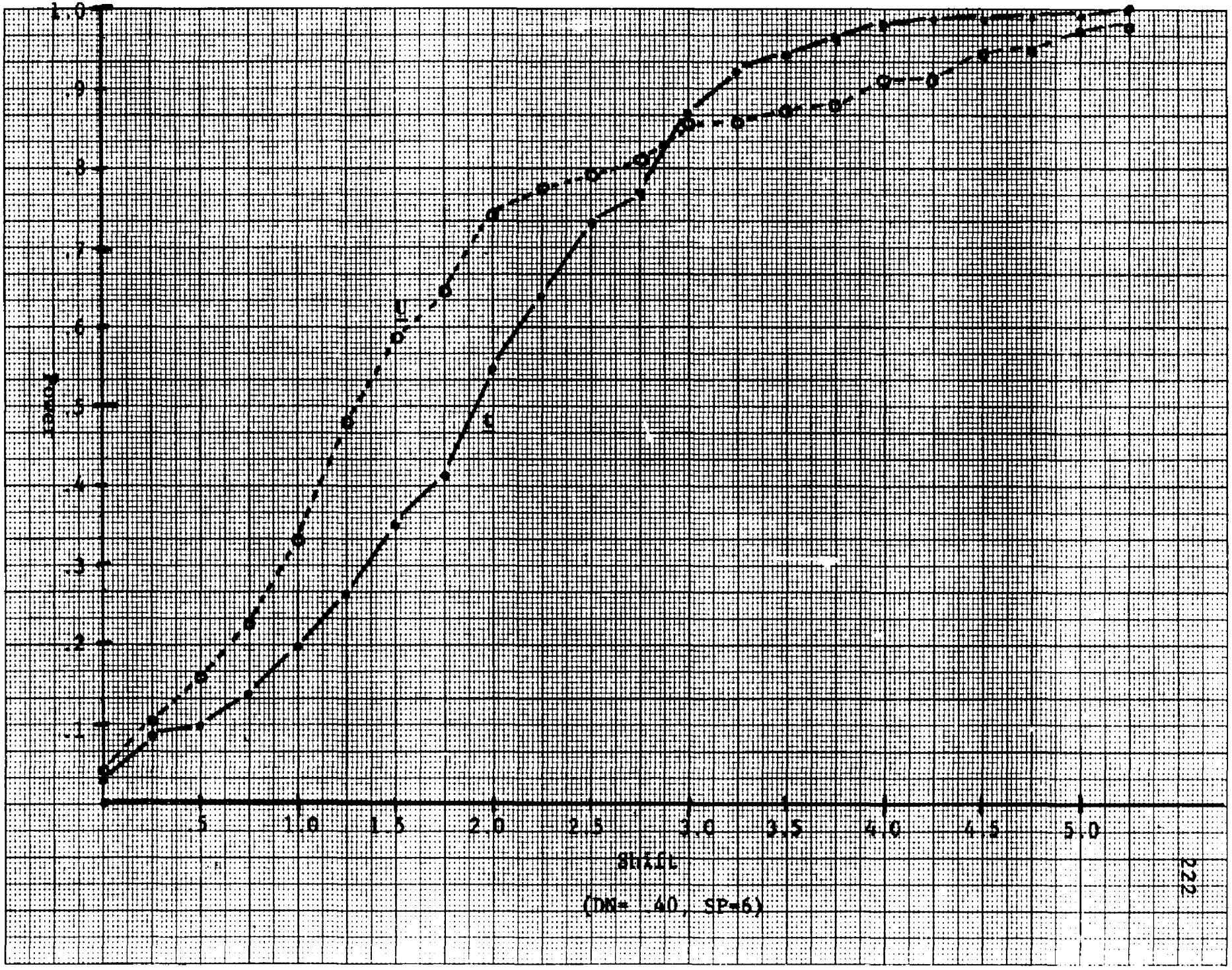
Millimeters to the Centimeter

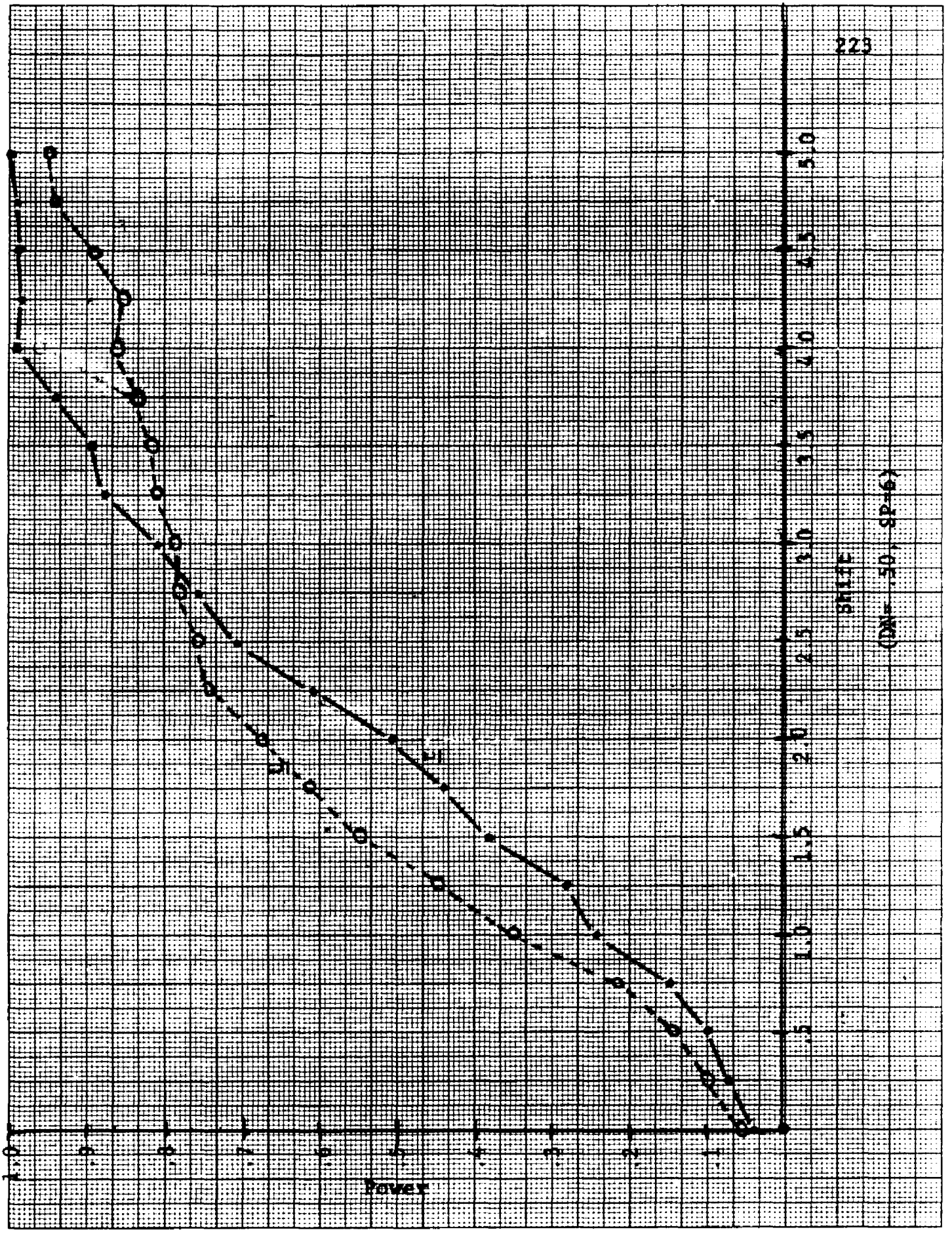


0 Millimeters to the Centimeter

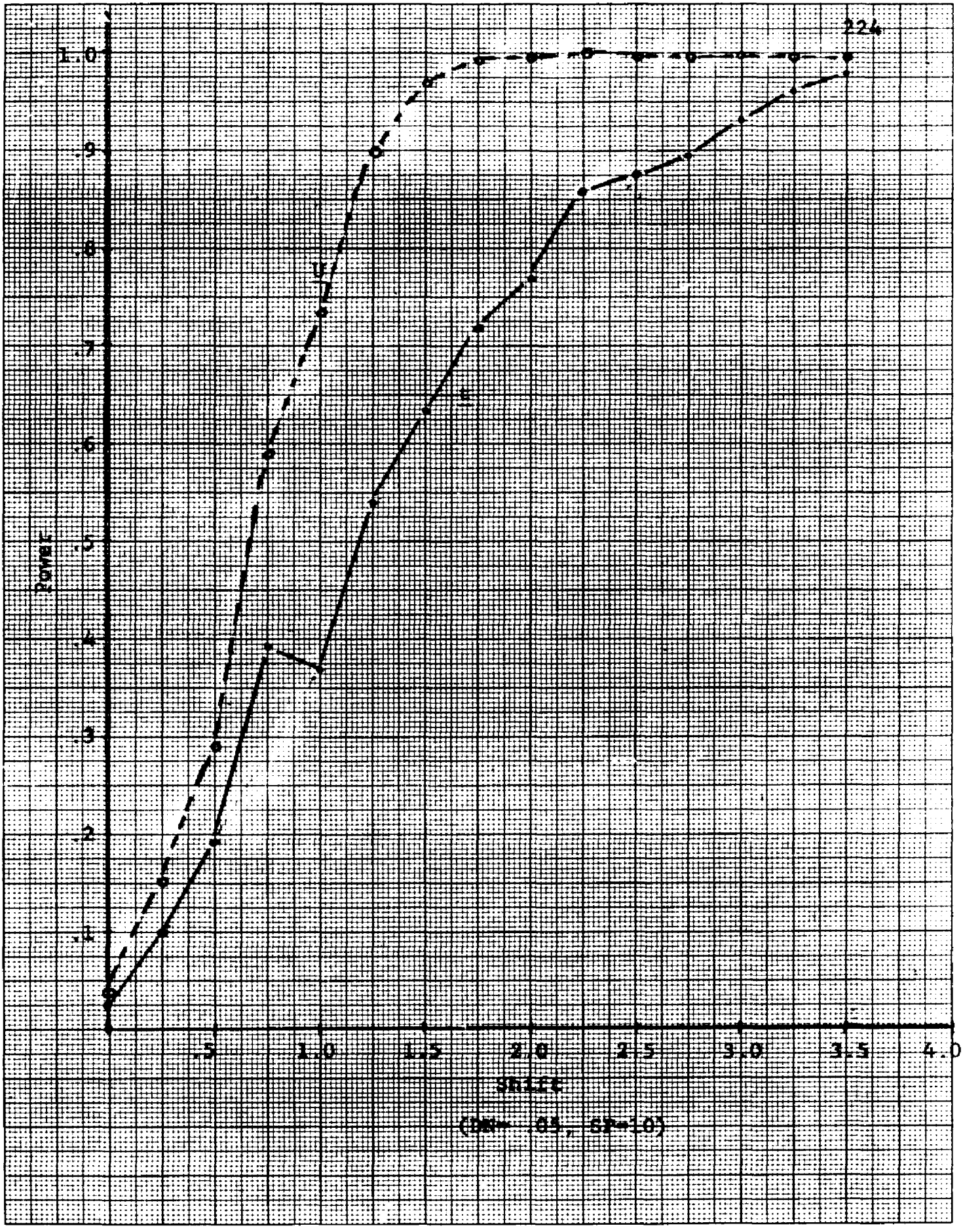


Millimeters to the Centimeter

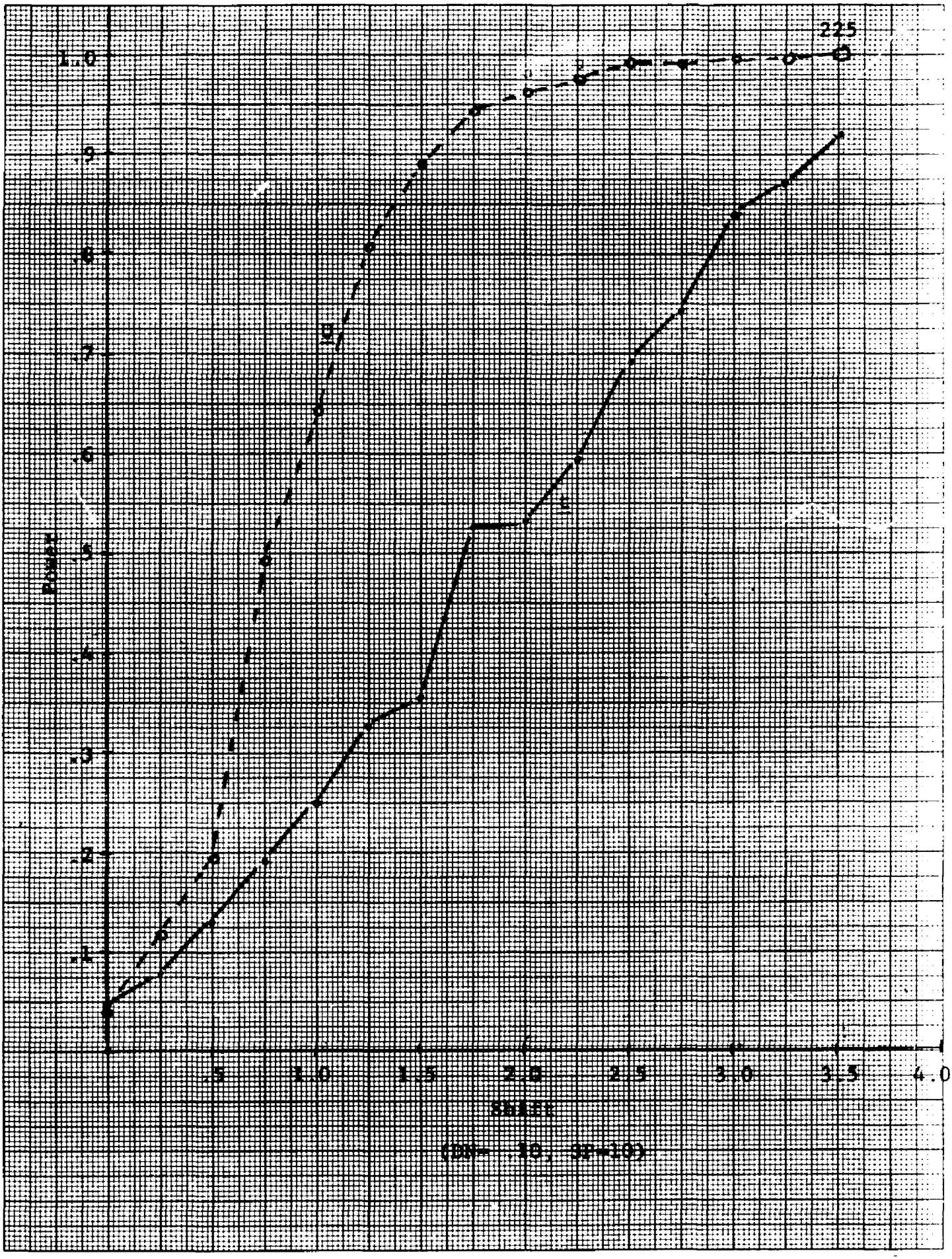

1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
force
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
shift
(DR= .40, SP=6)

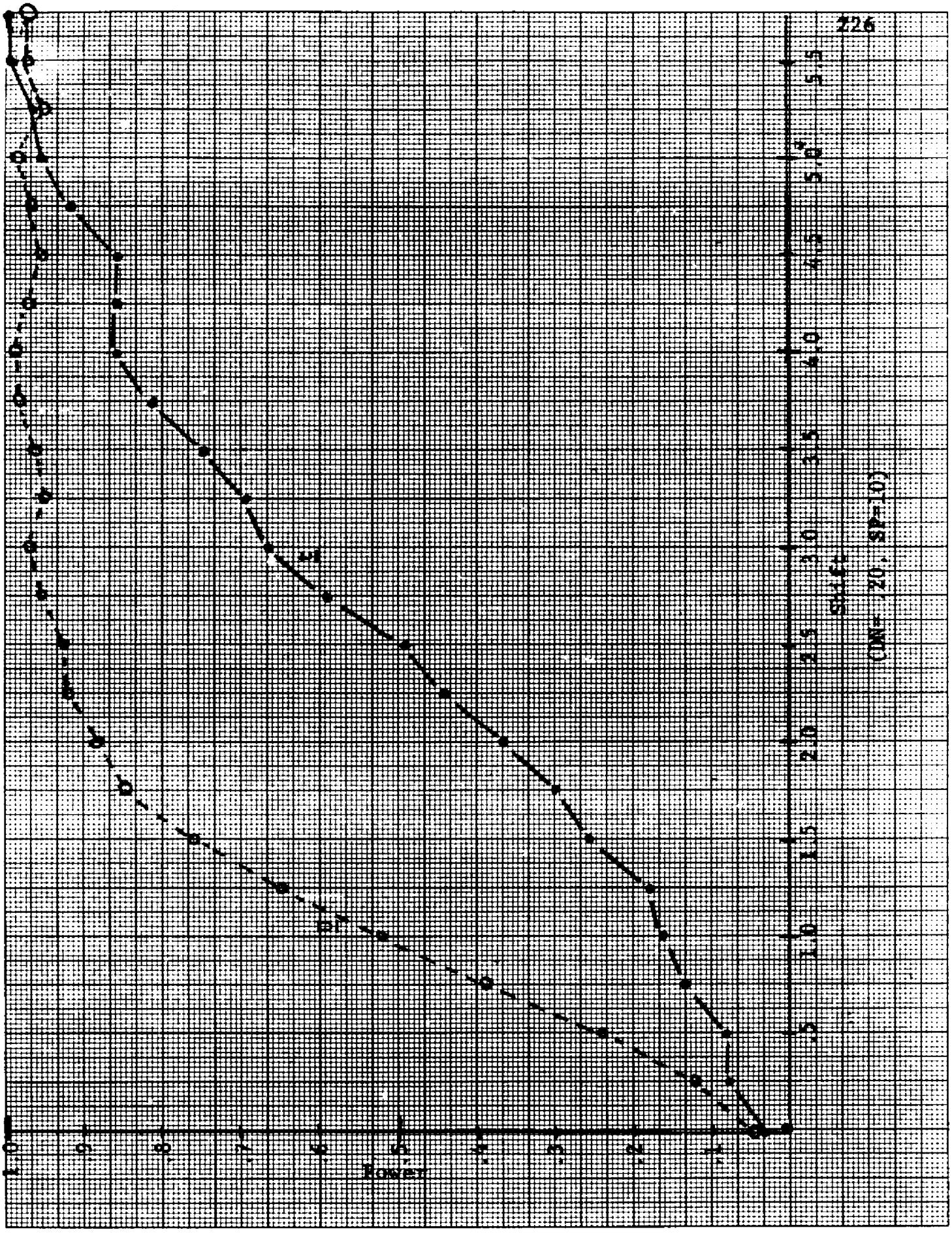


illimeters to the Centimeter

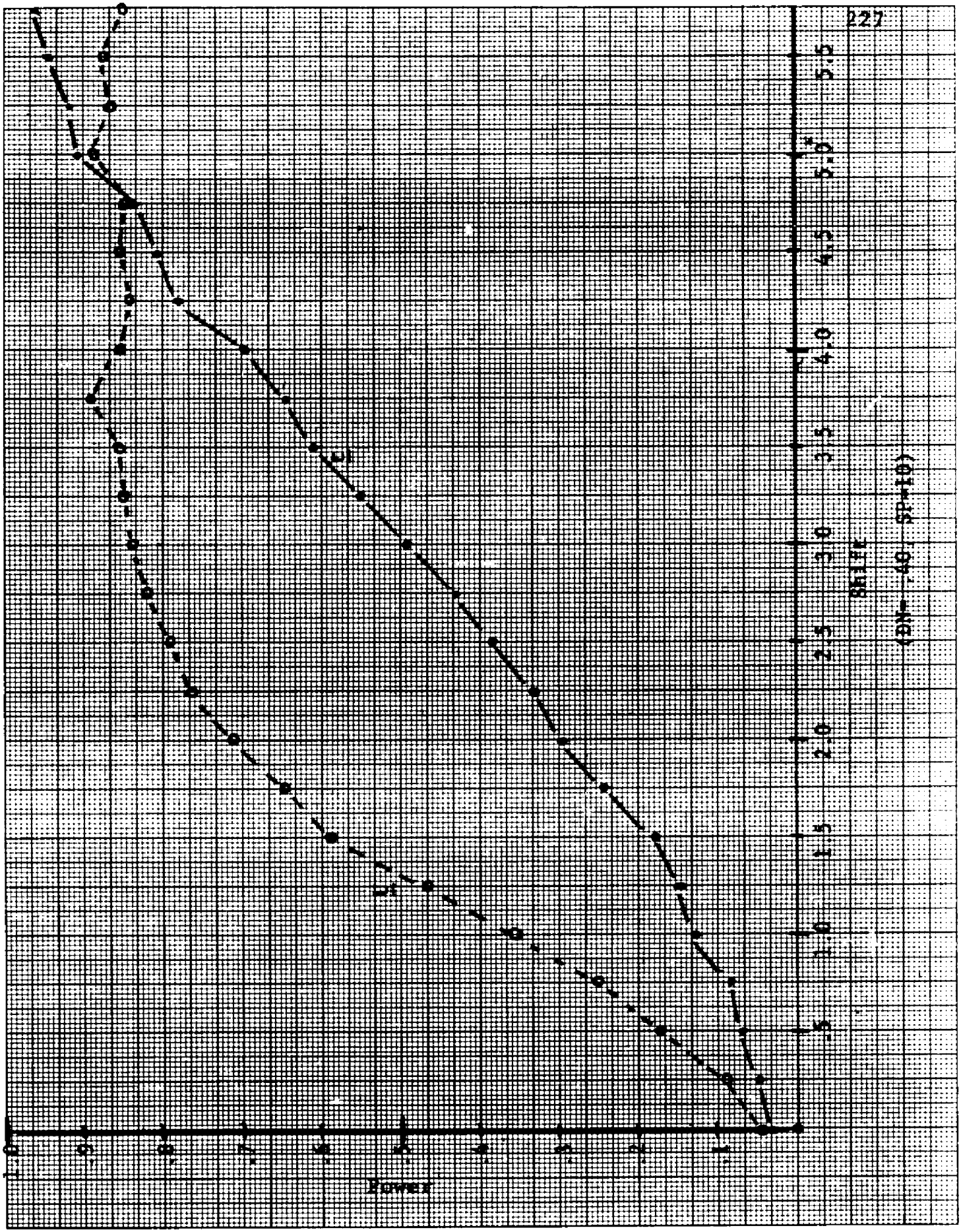


illimeters to the Centimeter

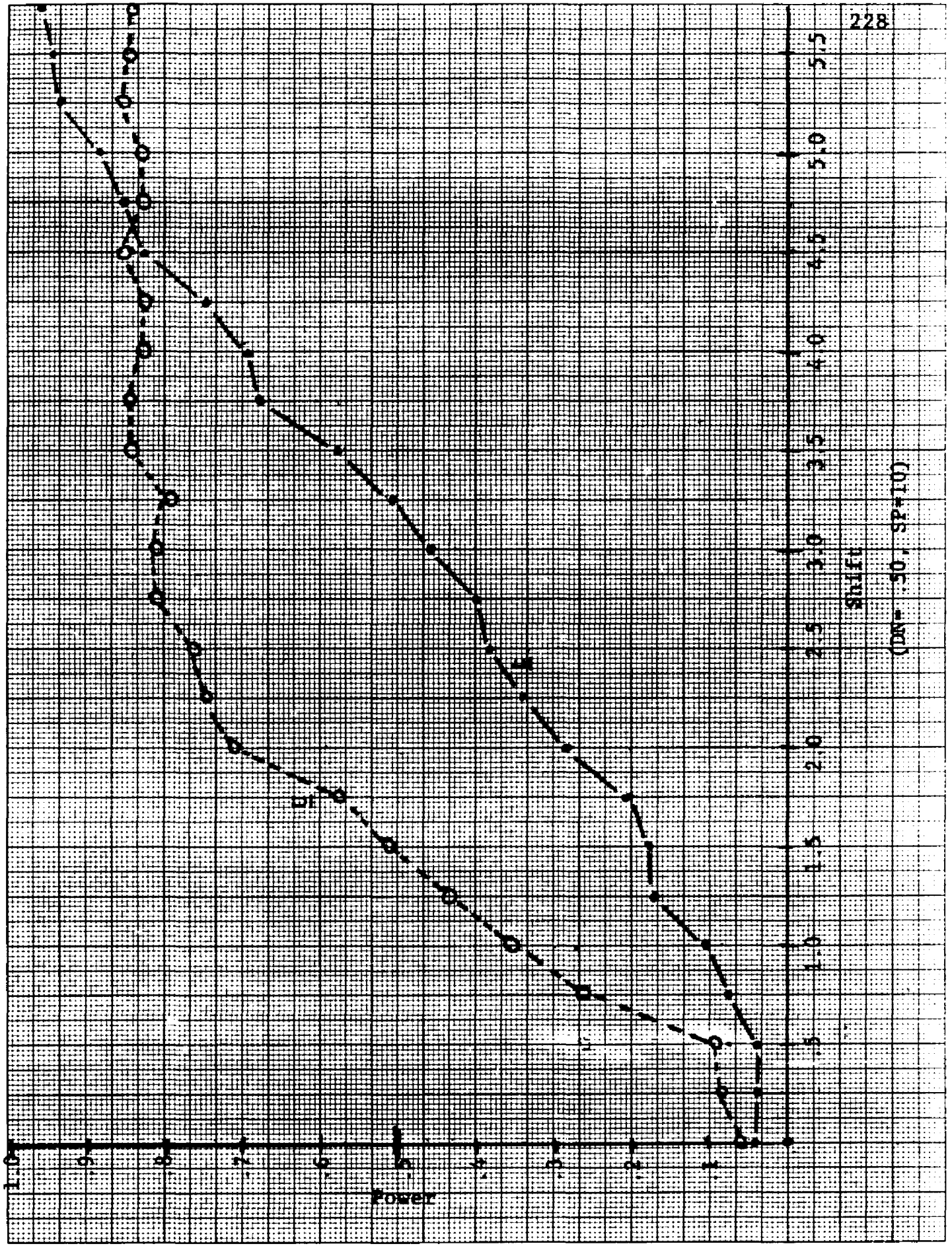




illimeters to the Centimeter



Millimeters to the Centimeter


Stiff
(DR= 50, SP=10)
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Power

BIBLIOGRAPHY

Allport, F. H. "The J-Curve Hypothesis of Conforming Behavior." Journal of Social Psychology 5 (1934): 141-83.

Blalock, Hubert M., Jr. Social Statistics. New York: McGraw-Hill, 1972.

Blair, R. Clifford and Higgins, James J. "A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's t Statistic Under Various Nonnormal Distributions." Journal of Educational Statistics 5, (1980): 309-35.

Blair, R. Clifford; Higgins, James J.; and Smitley, William D. S. "On the Relative Power of the U and t Tests." British Journal of Mathematical and Statistical Psychology 33 (1980): 114-20.

Boneau, C. Alan. "A Comparison of the Power of the U and t Tests." Psychological Review 69 (1962): 246-56.

Boneau, C. Alan. "The Effects of Violations of Assumptions Underlying the t-test." Psychological Bulletin 57 (1960): 49-64.

Box, Joan Fisher. "Gosset, Fisher, and the t Distribution." The American Statistician 35 (May 1981): 61-6.

Bradley, James V. "A Common Situation Conductive to Bizarre Distribution Shapes." The American Statistician 31 (1977): 147-50.

Bradley, James V. Distribution-Free Statistical Tests. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Bradley, James V. "Robustness?" British Journal of Mathematical and Statistical Psychology 31 (1978): 144-52.

Conover, W. J. Practical Nonparametric Statistics. New York: John Wiley, 1971.

Dixon, W. J. "Power Under Normality of Several Nonparametric Tests." The Annals of Mathematical Statistics 25 (1959): 610-14.

Geary, R. D. "Testing for Normality." Biometrika 34 (1947): 209-42.

Glass, G. V.; Peckham, P. D.; and Sanders, J. R. "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance." Review of Educational Research 42 (1972): 237-88.

Hamburg, Morris: Statistical Analysis for Decision Making. New York: Harcourt, Brace, and World, Inc., 1970.

Hayman, G. E. and Govindarajula, Z. "Exact Power of the Mann-Whitney Test for Exponential and Rectangular Alternatives." The Annals of Mathematical Statistics 37 (1966): 945-53.

Hodes, J. L. and Lehmann, E. L. "The Efficiency of Some Non-parametric Competitors to the t-test." The Annals of Mathematical Statistics 27 (1956): 324-35.

Hotelling, H. and Pabst, M. R. "Rank Correlations and Tests of Significance Involving No Assumptions of Normality." The Annals of Mathematical Statistics 7 (1936): 29-43.

Kerlinger, Fred N. Foundations of Behavioral Research. New York: Holt, Rinehart, and Winston, 1973.

Kleijnen, Jack P. C. Statistical Techniques in Simulation, Part 1. New York: Marcel Dekker, Inc., 1974.

Lehman, E. L. Nonparametrics. San Francisco: Holden-Day, 1975.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton, Mifflin, 1953.

Mann, H. B. and Whitney, D. R. "On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other." The Annals of Mathematical Statistics 18 (1947): 50-60.

Mikulski, Piotr W. "On the Efficiency of Optimal Nonparametric Procedures in the Two Sample Case." The Annals of Mathematical Statistics 34 (1963): 22-32.

Mood, A. M. "On the Asymptotic Efficiency of Certain Non-Parametric Two-Sample Tests." The Annals of Mathematical Statistics 25 (1954): 514-22.

Neave, Henry and Granger, C. W. J. "A Monte Carlo Study Comparing Various Two Sample Tests for Differences in Mean." Technometrics 10 (1968): 509-22.

Nie, Norman H., etc. al. SPSS: Statistical Package for the Social Sciences, 2nd ed. New York: McGraw-Hill, 1975.

Noether, Gottfrield E. "On a Theorem of Pitman." The Annals of Mathematical Statistics 26 (1955): 64-8.

Pearson, E. S. "The Analysis of Variance in Cases of Non-Normal Variation." Biometrika 23 (1931): 114-33.

Runyon, Richard P. and Haber, Audrey. Fundamentals of Behavioral Statistics. Reading, Massachusetts: Addison-Wesley Publishing Co., 1976.

Savage, I. Richard. "Bibliography of Nonparametric Statistics and Related Topics." American Statistical Association Journal 48 (1953): 844-906.

Scheffé, Henry. The Analysis of Variance. New York: John Wiley, 1959.

Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

Stone, M. "Extreme Tail Probabilities for the Null Distribution of the Two-Sample Wilcoxon Statistic." Biometrika 54 (1967): 629-40.

"Student." "The Probable Error of a Mean." Biometrika 6 (1908): 1.

Ulam, Stanislaw. Mathematical Thinking in the Behavioral Sciences. New York: W. C. Freeman, 1968.

White, C. "The Use of Ranks in a Test of Significance for Comparing Two Treatments." Biometrics 8 (1952): 33-41.

Wilcoxon, F. "Individual Comparisons by Ranking Methods." Biometrics 1 (1945): 80-3.

Zuwaylif, Fadil H. Applied Business Statistics. Reading, Massachusetts: Addison-Wesley, 1974.

State University of Iowa. Norton, D. W., "An Empirical Investigation of Some Effects of Non-Normality and Heterogeneity of the F-Distribution," 1952.

ABSTRACT

A MONTE-CARLO INVESTIGATION OF STATISTICAL POWER UNDER THE MIXED-NORMAL DISTRIBUTION

by

Walter H. Mackey II

August, 1982

Adviser: Dr. Joseph Posch, Jr.

Major: Educational Evaluation and Research

Computer simulations were used to compare the power and the Type-I error rates for the "Student" t-test and the Mann-Whitney U-Test. Sample sizes were eighteen. These samples were randomly drawn from population frequency distributions that were mixed-normal, a distribution form in which the total population was composed of two separate sub-populations, each of which was normally distributed. Several mixed-normal distributions were examined for different values of Density (DN), the proportion of cases in one sub-population, Separation (SP), the distance between the two sub-population means, and sub-population standard deviation ratios (DEVRATIO). The research hypotheses were directional, and the alpha-level was 0.05.

It was concluded that 1) both tests maintained Type-I error rates that were generally consistent with the theoretical value of 0.05 and that 2) the Separation Factor was the

most critical factor in judging power superiority. For small measures of Separation (two or less standard units), the t-test held a slight power advantage. At the larger values of Separation (four or more standard units), the U-Test was generally the more powerful test. Whenever a large Separation Factor was coupled with a Density value greater than or equal to 0.20, the U-Test maintained a power advantage for small measures of Shift (difference between Experimental and Control group means). As the Shift value increased, however, the t-test regained power advantage under certain mixed-normal distributions. It was impossible to claim a power superiority for either test under these conditions. DEVRATIO had little effect on the relative power between the two tests.

Unless a large Separation Factor (greater than two standard units) exists in the population that is being tested, the choice of the "Student" t-test would be most correct under mixed-normality. For larger Separation values, the U-Test has a lower probability of Type-II error.

AUTOBIOGRAPHICAL STATEMENT

Name: Walter Henry Mackey II

Date of Birth: August 30, 1947

Education: Detroit St. Charles High School (1965)

Wayne State University, Bachelor of Arts with Distinction (1969)

University of Detroit, Master of Arts in Teaching Mathematics (1972)

Wayne State University, Doctor of Philosophy in Evaluation and Research (1982)

Education Work Experience:

Detroit St. Charles Junior High School (1969-71)

Detroit St. Rita High School (1971-72)

Grosse Pointe South High School (1972-present)

Memberships:

National Council of Teachers of Mathematics

Detroit Area Council of Teachers of Mathematics

Michigan Council of Teachers of Mathematics

Phi Beta Kappa, Michigan Gamma Chapter

American Educational Research Association

American Statistical Association