CONCENTRATION BANDS FOR UNIFORMITY PLOTS

C. P. Quesenberry and Craige Hales

North Carolina State University
Raleigh, North Carolina 27650, USA

A graphical method for deciding whether a set of numbers may be considered to be observed values of a random sample on a uniform random variable on the unit interval consists of plotting the ranked values against the expected values of uniform order statistics. It is suggested here that curves called concentration curves drawn on the same or supplementary graphs are helpful in assessing the uniformity of the plotted points. Prototype graphs are given for two examples, and fifteen graphs are given for selected sample sizes that may be used to supplement data plots.

1. INTRODUCTION AND SUMMARY

In recent papers Quesenberry and Miller (1977), QM, and Miller and Quesenberry (1979), MQ, have studied power properties of many goodness-of-fit tests for testing that a set of values are observations on a set of independent uniform random variables on the unit interval. The motivation for interest in testing uniformity stems from the classical probability integral transformation, and from recent work on conditional probability integral transformations by the first author of this paper and co-workers. This work includes papers by O'Reilly and Quesenberry (1973), Quesenberry (1975), Quesenberry and Starbuck (1976), Quesenberry, Whitaker and Dickens (1976), and Rincon-Gallardo, Quesenberry, and O'Reilly (1979).

For a general discussion of the problem of testing uniformity and references to much of the literature in this area the reader should see the papers by QM and MQ.

Aside from formal goodness-of-fit tests for uniformity, it is often helpful in studying the uniformity of a set of numbers to make graphical plots of

the data. One particular type of plot that is often helpful is to plot the ordered values of the numbers against the expected values of the uniform order statistics, as suggested in Quesenberry, et al (1976) and in Quesenberry (1979). If the numbers are, in fact, observed values of a sample from a uniform distribution on the unit interval, then the plotted points should follow the straight line connecting the points (0, 0) and (1, 1). The question is, how closely should the plotted points follow the line under the uniformity assumption? The procedure, while useful, is rather subjective and we feel that supplementary graphs that give information about the concentration of the individual order statistics about their expected values are helpful in interpreting these uniformity plots.

2. NOTATION AND DESCRIPTION OF CONCENTRATION CURVES

Denote by $u_1, \cdots, u_N$ a set of N independently and identically distributed uniform random variables on the unit interval (0, 1). In this paper we shall use lower case u's for both these random variables and for observations on them. For $u_{(1)} \leq \cdots \leq u_{(N)}$ the order statistics for this sample, then $u_{(j)}$ is a beta random variable with parameters (j, N - j + 1), mean $[j/(N+1)]$, and variance $[j(N-j+1)/(N+1)^2(N+2)]$. A plot of the points $(u_{(j)}, j/(N+1))$ for $j = 1, \cdots, N$, should follow the line $g(u) = u$ for $0 < u < 1$. How closely the points should follow the line depends, of course, on the sample size N, so the question arises as to how closely they should follow it for particular values of N.

Let $I_x(a, b)$ denote the incomplete beta function, i.e.,

$$I_x(a, b) = \int_0^x y^{a-1}(1-y)^{b-1}dy \Big/ \int_0^1 y^{a-1}(1-y)^{b-1}dy \ ,$$

for $0 < x < 1$, $a > 0$, and $b > 0$. Also, let $I_p^-(a, b)$ denote the inverse beta function, i.e.,

$$x = I_p^-(a, b) \text{ iff } I_x(a, b) = p, \ 0 < p < 1 \ .$$

In Figures 3-17 of Section 4 for each value of N in the set {2, 5, 10, 15, 20, 30, 40, 50, 60, 80, 100, 150, 200, 300, 500}, we have computed, plotted and drawn smooth curves connecting the N points $[I_p^-(j, N-j+1), j/(N+1)]$; $j = 1, \cdots, N$; for p assuming each of the values in the set {.01, .05, .20, .35, .65, .80, .95, .99}. This gives a picture of the concentration of probability about the line $g(u) = u$, $0 < u < 1$, which is convenient for reference in interpreting probability plots. We shall call these curves concentration curves, and the area between the two corresponding curves on each side of the line $g(u) = u$ a concentration band. For example, the two outside curves enclose a 98% concentration band, the next two give a 90% concentration band, etc.

Example 2.1 The observed values of a set of N = 10 observations were as follows: 0.20, 0.23, 0.35, 0.40, 0.41, 0.45, 0.50, 0.72, 0.84, 0.91. Figure 1 shows a plot of these ten values as well as the 98, 90, 60, and 30 per cent concentration bands. These values are seen in Figure 1 to be very uniform.

We have also computed two statistics which are good measures of uniformity. The modified Watson statistic, U^2_{MOD}, (cf. Stephens (1970), QM) here has the value $U^2_{MOD} = 0.084$, which is much smaller than the upper 10 per cent point which is 0.151 (cf. QM, p.173). The Neyman (1937) smooth statistic, p_4^2, here has the value $p_4^2 = 2.83$ which is also much smaller than its upper 10 per cent point which is here 9.64 (cf. MQ, p. 276). These statistics thus support the uniformity conclusion arrived at from perusal of Figure 1.

Example 2.2 A sample of size N = 30 gave the values:

0.023	0.032	0.054	0.069	0.081	0.094
0.105	0.127	0.148	0.169	0.188	0.216
0.255	0.277	0.311	0.361	0.376	0.395
0.432	0.463	0.481	0.519	0.529	0.567
0.642	0.674	0.752	0.823	0.887	0.926

Figure 2 shows a plot of these values against their expected values and the 98, 90, 60 and 30 per cent concentration bands. We feel that Figure 2 casts considerable doubt on a uniformity assumption for this data since the trail of plotted points lies consistently outside the 90% concentration band over much of the range of the u-values.

The values of the U^2_{MOD} and p^2_4 statistics have been computed for this data and are $U^2_{MOD} = 0.116$ and $p^2_4 = 6.62$. This value of U^2_{MOD} is to be compared with the upper 10% point of 0.151 and while it is not significant, it is larger than for the first example. The p^2_4 statistic is approximately a $\chi^2(4)$ random variable under uniformity, and so the significance level observed here is $P[\chi^2(4) > 6.62] = 0.16$.

3. REMARKS ON USAGE

The reader who wishes to use concentration bands with uniformity plots as in Figure 1 and 2 may proceed to do so in a number of ways. Some users will have access to computers that have available inverse incomplete beta function subroutines and peripheral plotting equipment, and can readily write a program to both plot the data points and the concentration curves on the same graph as in Figure 1 and 2. Of course, they may choose values of p other than those we have chosen here.

If the data are plotted by band, or otherwise, on standard graph paper then the graphs may be visually compared with the concentration curves in Figures 3, ···, 17. The comparisons will be especially easy if the plotting is done on tracing paper, or a transparency, and used as an overlay on the figures given here. The same scale must be used as for the figures here, of course. Also, we give here the concentration curves for only selected values of N. These values were chosen so that for each increasing value the concentration curves are slightly, but noticeably, tighter about the line $g(u) = u$. If it is

desired to plot a sample for a value of N for which a graph is not given here, then the plot for this value of N may be compared with the concentration curves for the smallest given value of N that exceeds the sample size in band. This rule results in a procedure which is slightly conservative in that it will tend to make the sample appear to be slightly less uniform than it, in fact, is.

4. FIGURES 1-17

Each of the graphs in the following figures gives 98, 90, 60 and 30 per cent concentration bands for the sample sizes indicated.

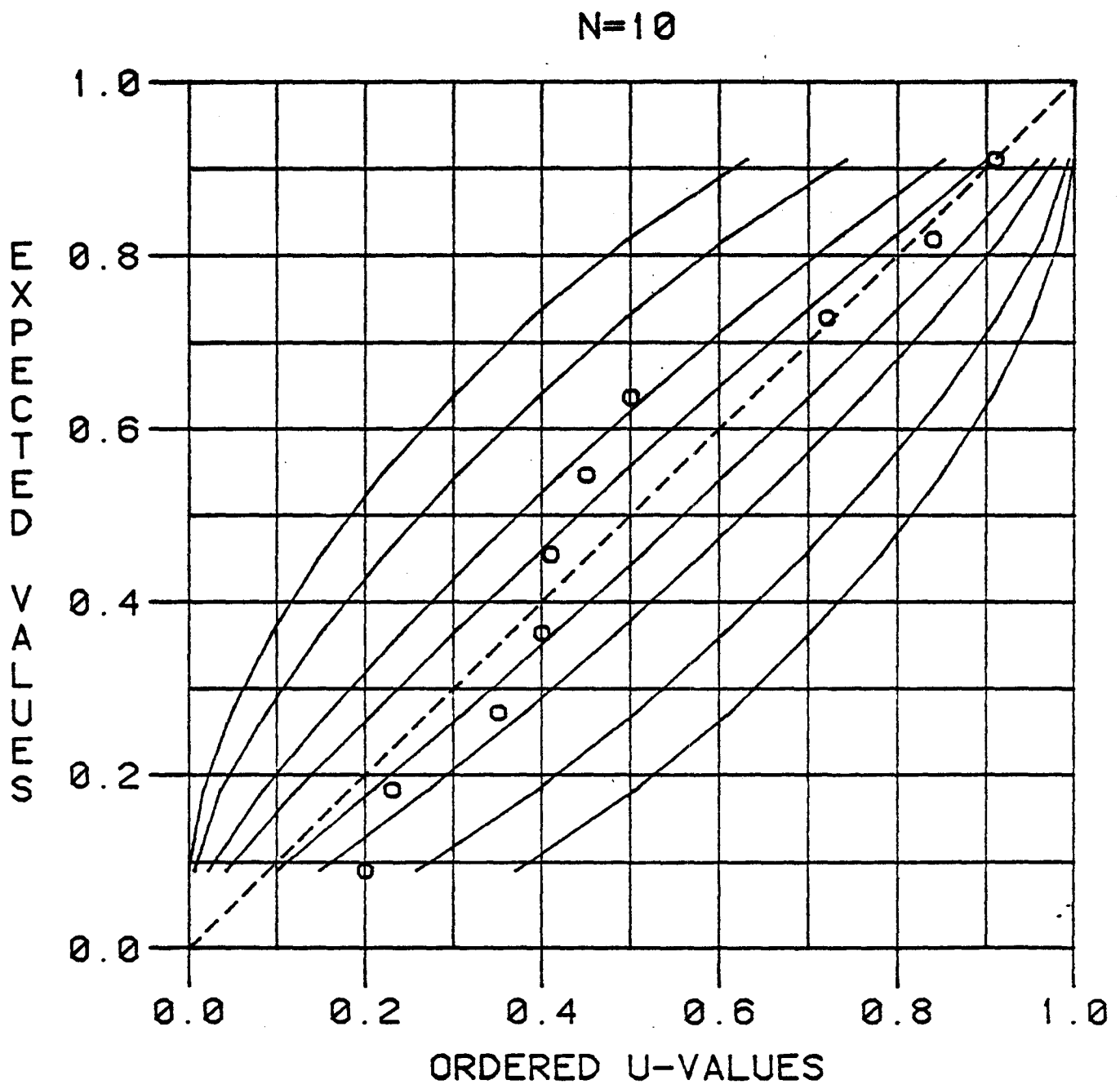

Figure 1: Graph for Example 2.1

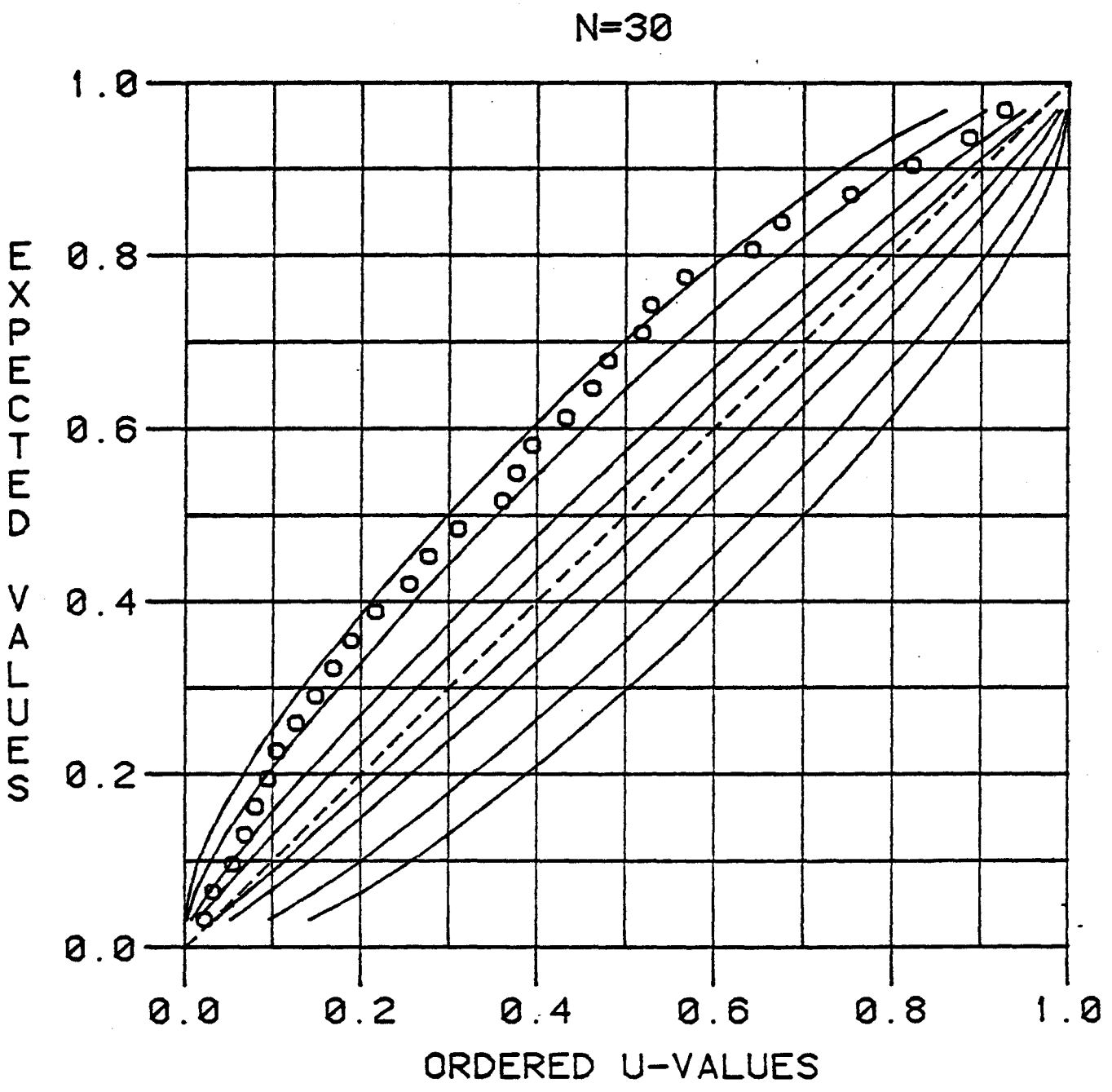

Figure 2: Graph for Example 2.2

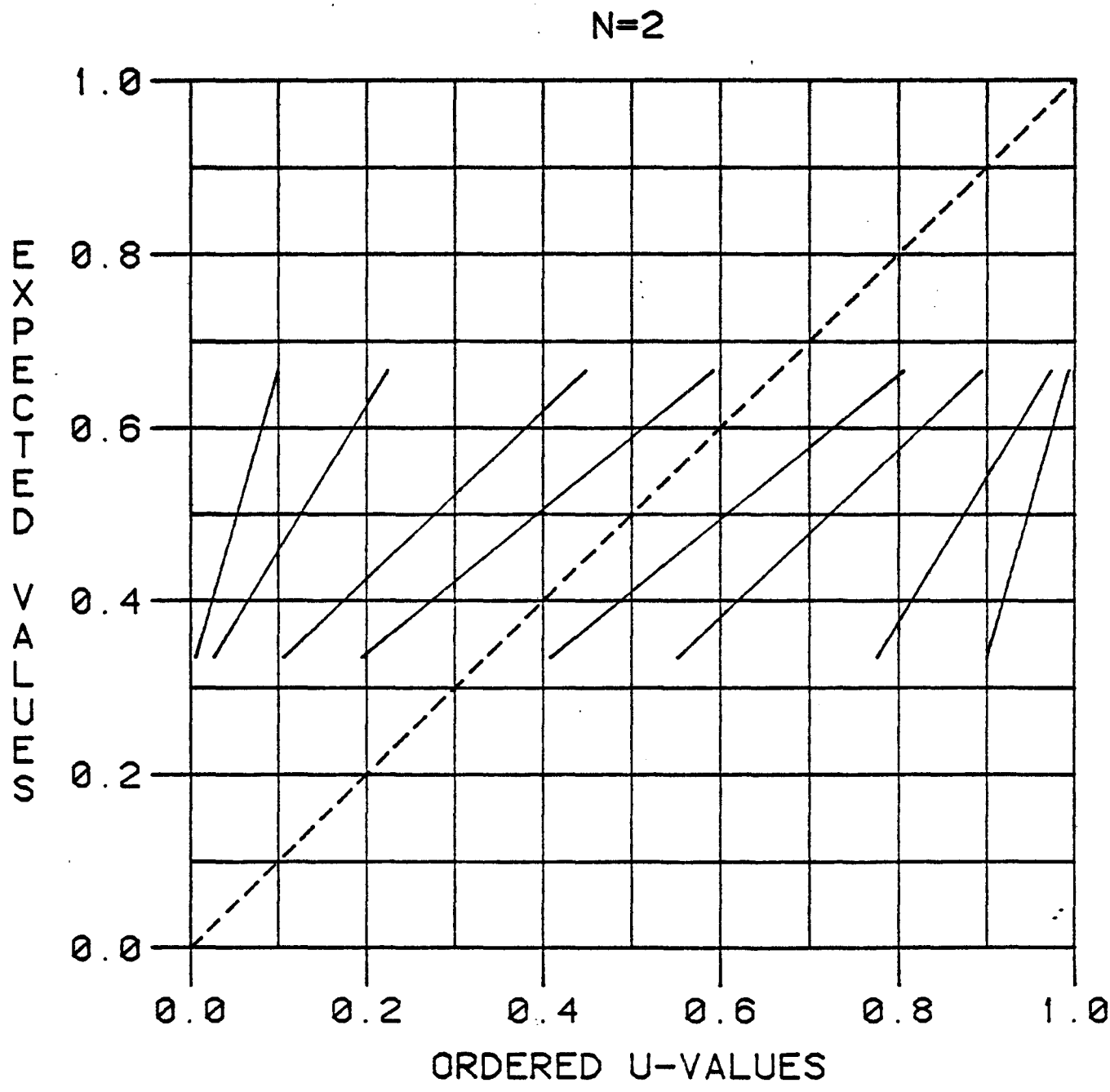

Figure 3: Concentration Bands for N = 2

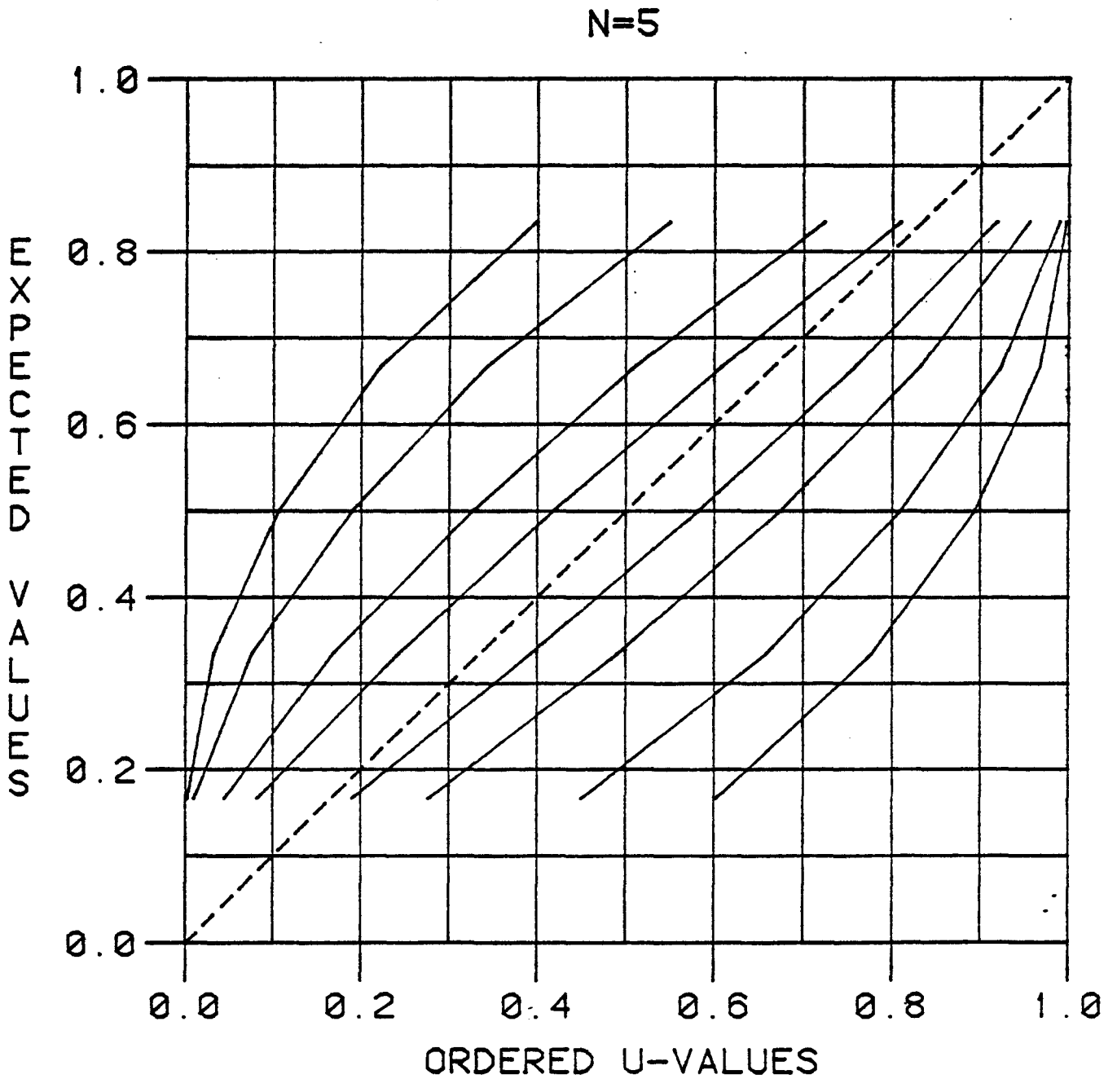

Figure 4: Concentration Bands for N = 5

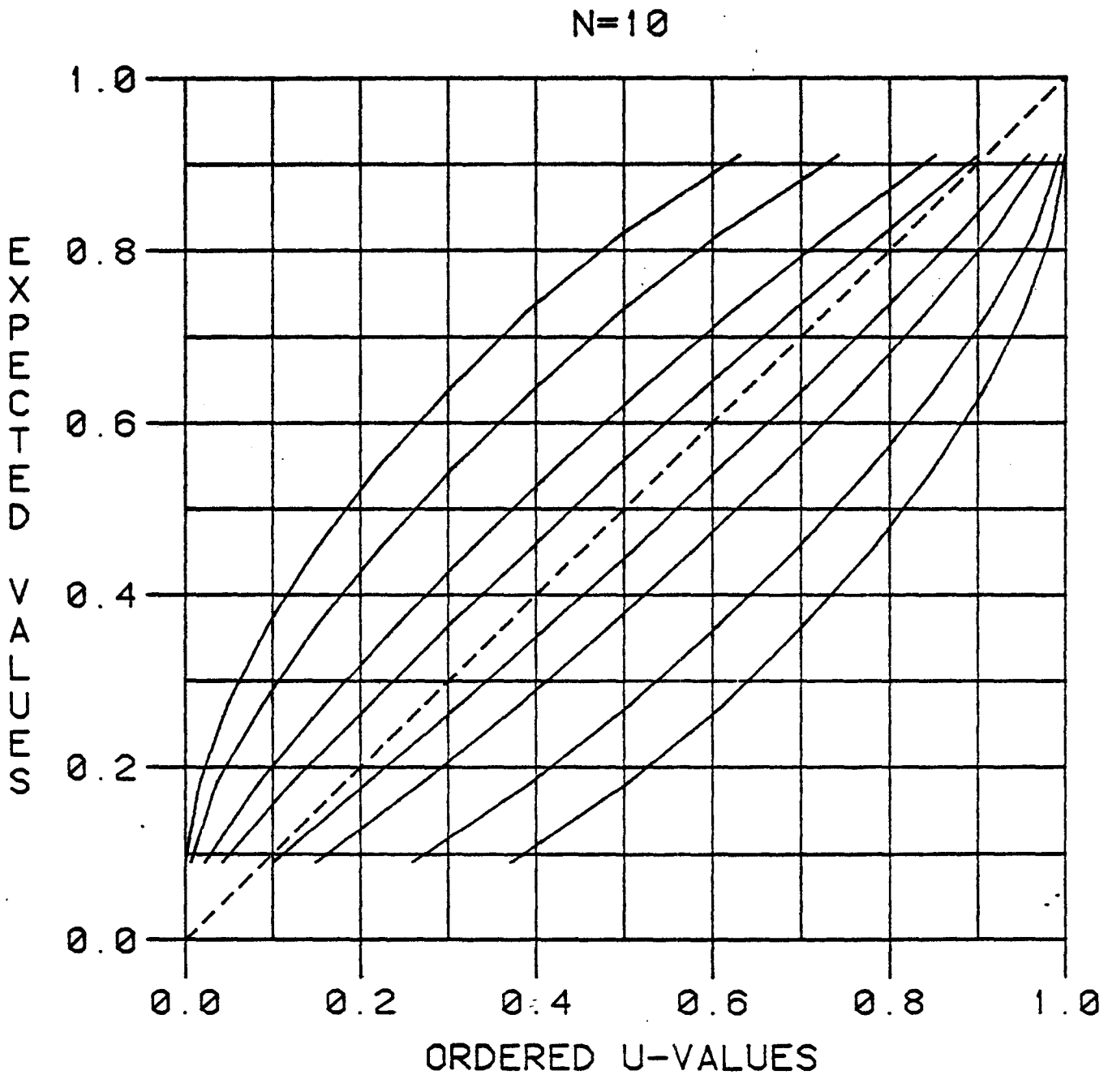

Figure 5: Concentration Bands for N = 10

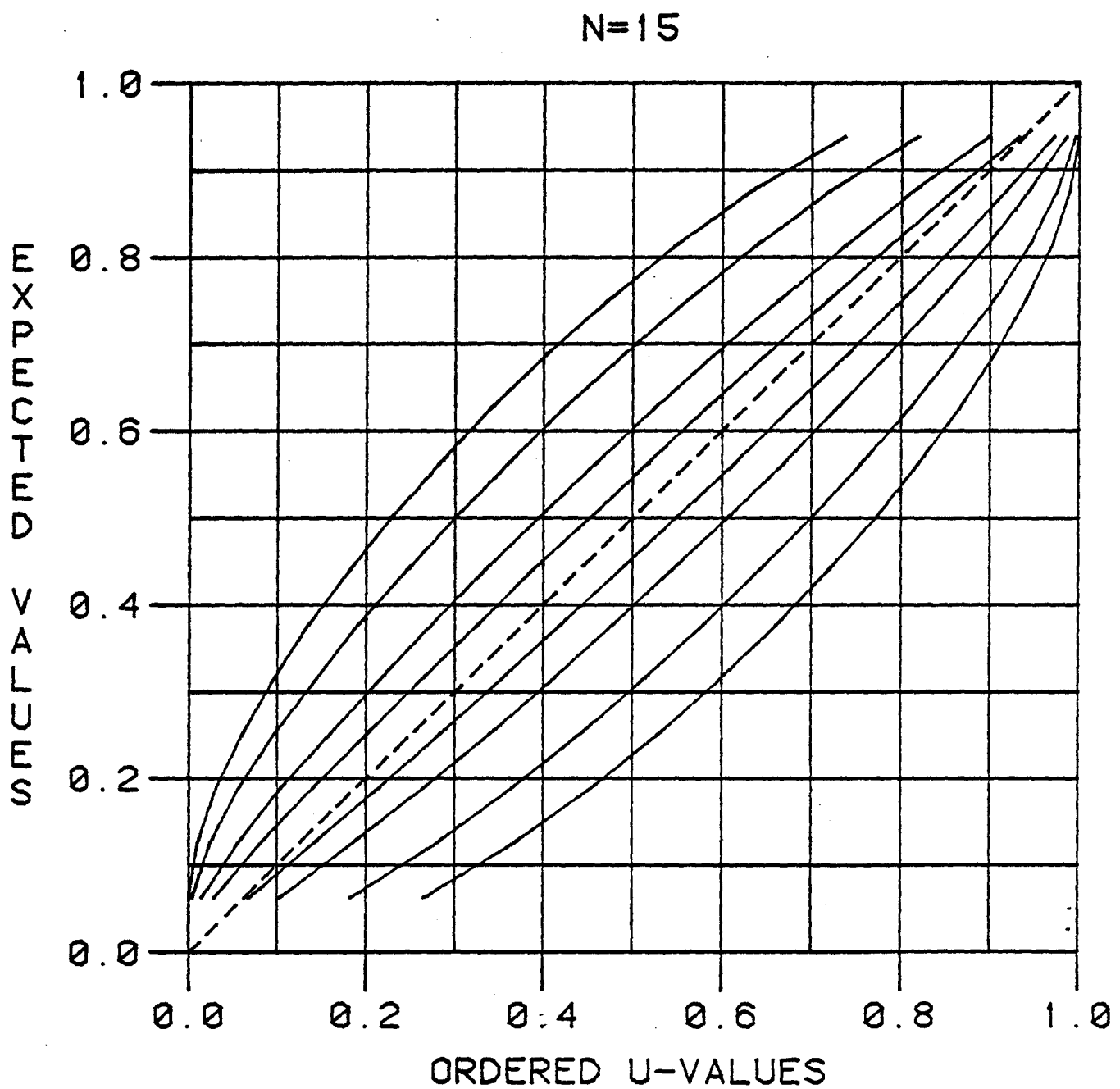

Figure 6: Concentration Bands for N = 15

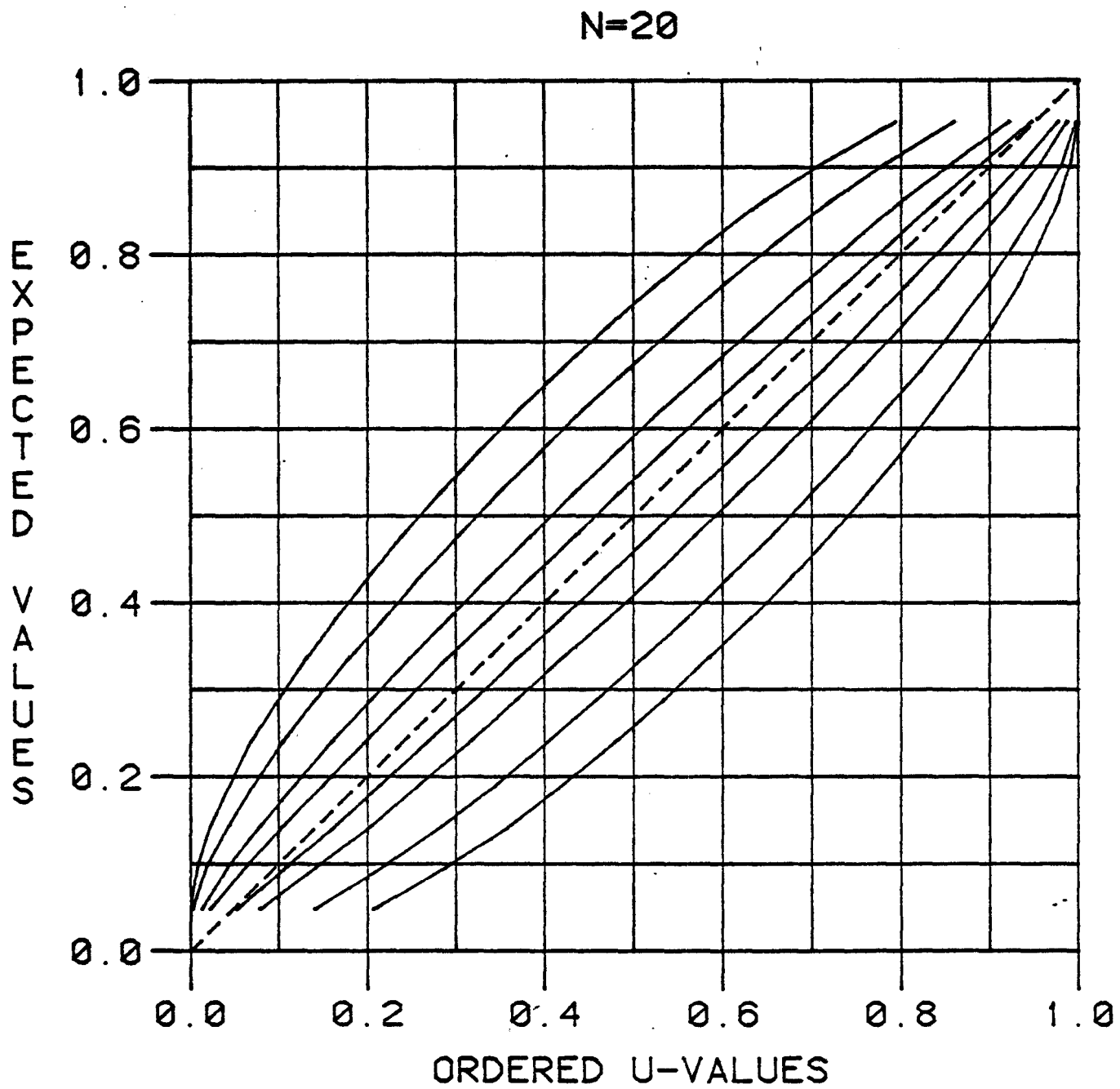

Figure 7: Concentration Bands for N = 20

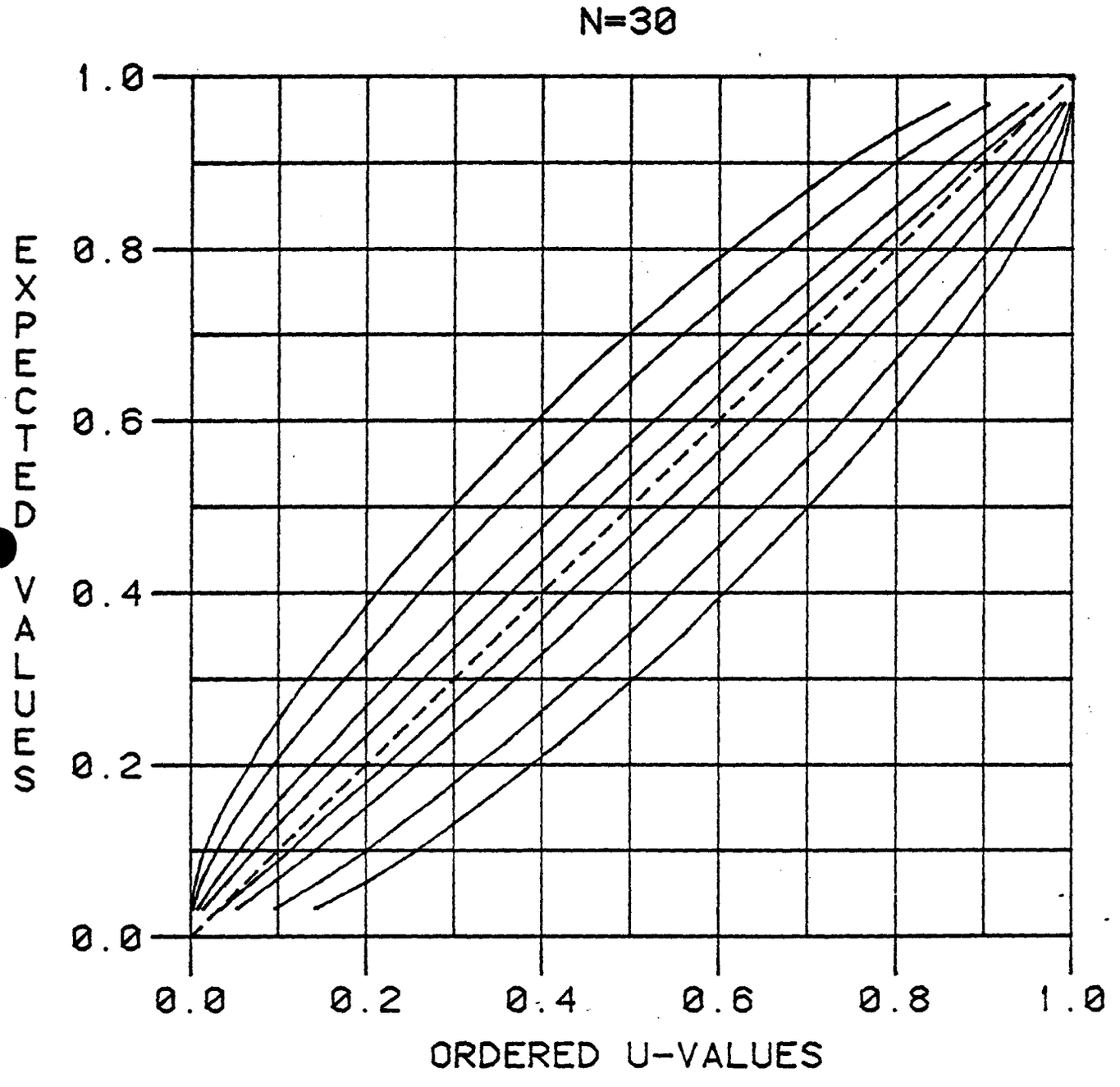

Figure 8: Concentration Bands for N = 30

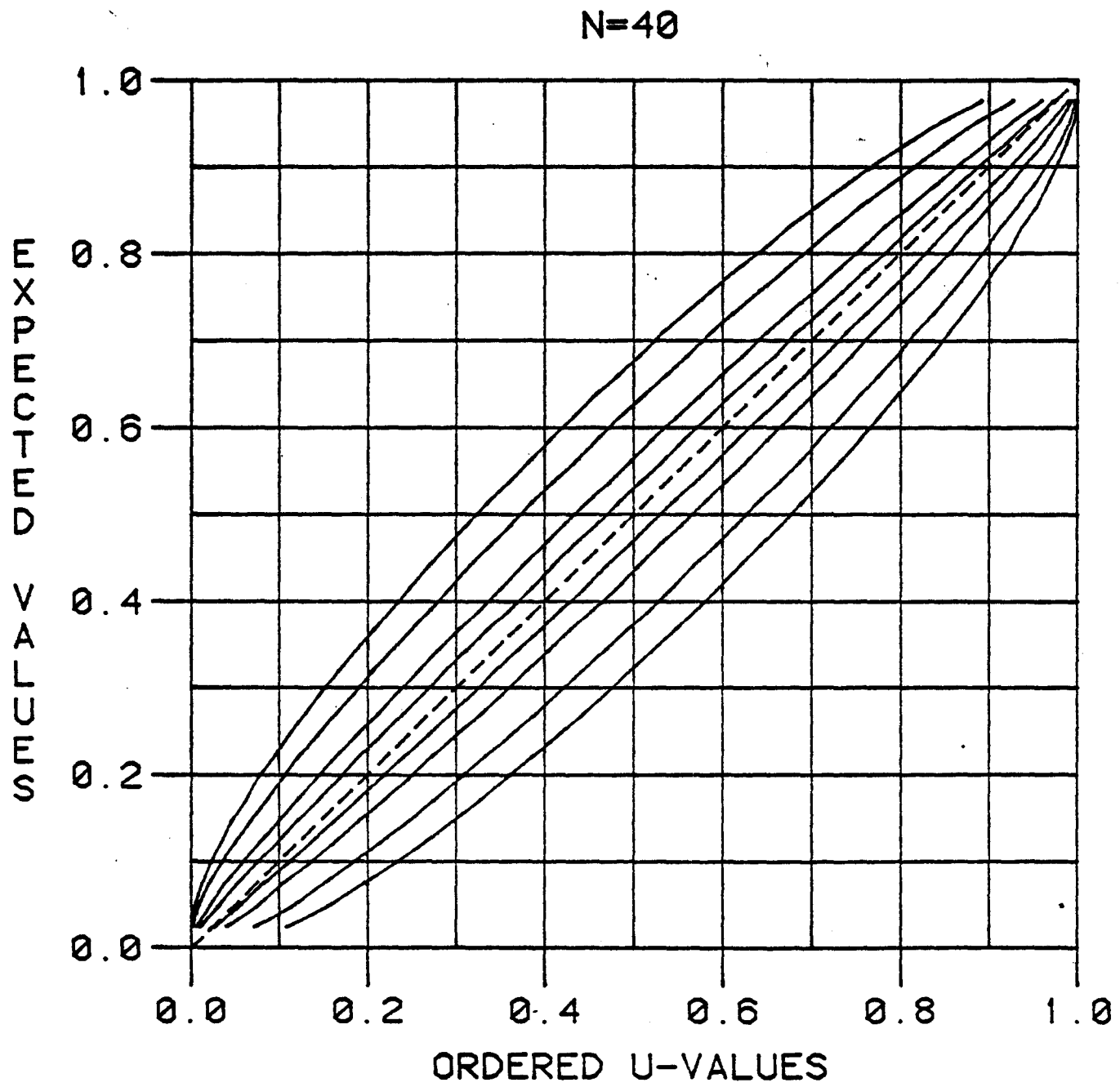

Figure 9: Concentration Bands for N = 40

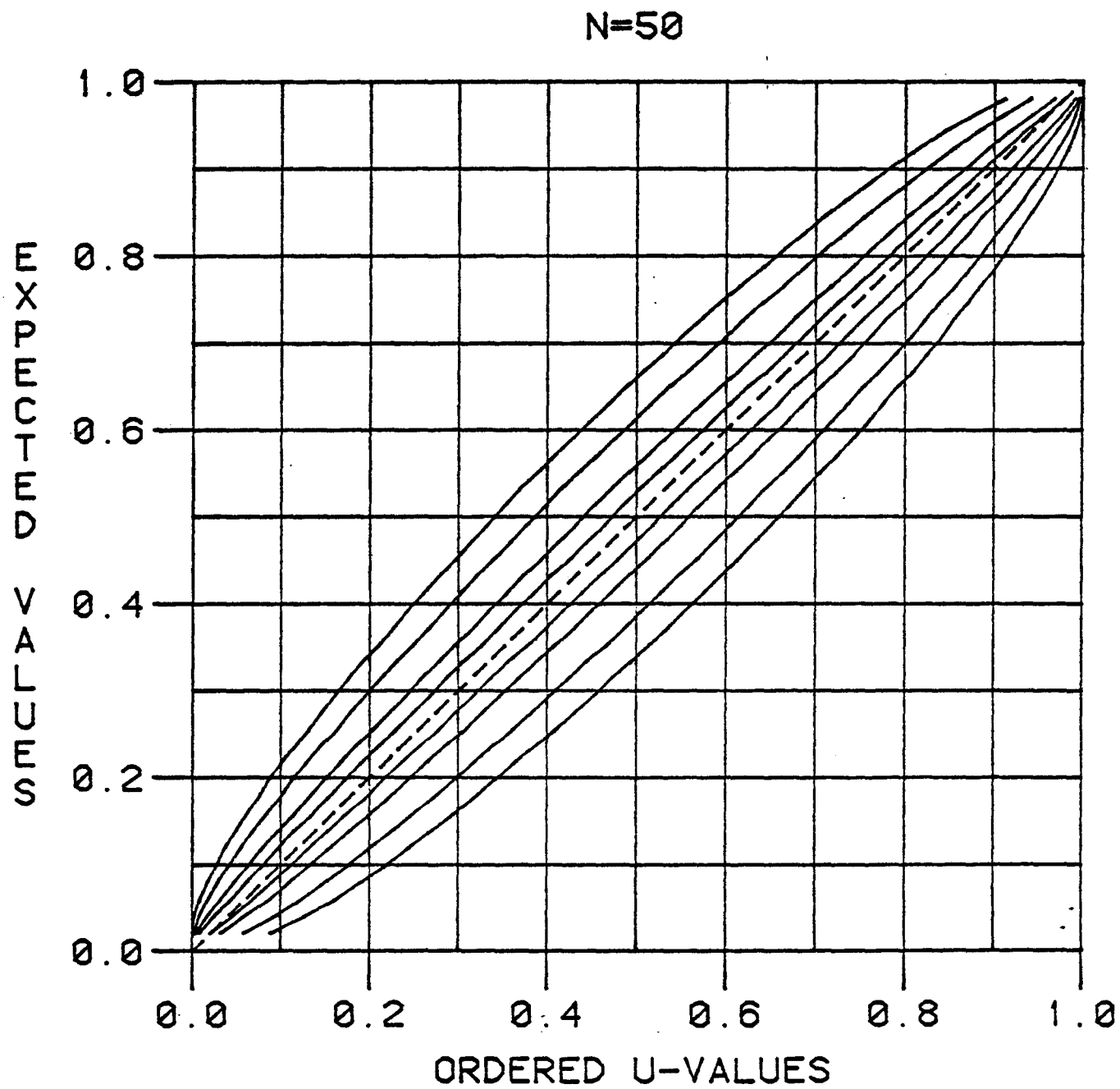

Figure 10: Concentration Bands for N = 50

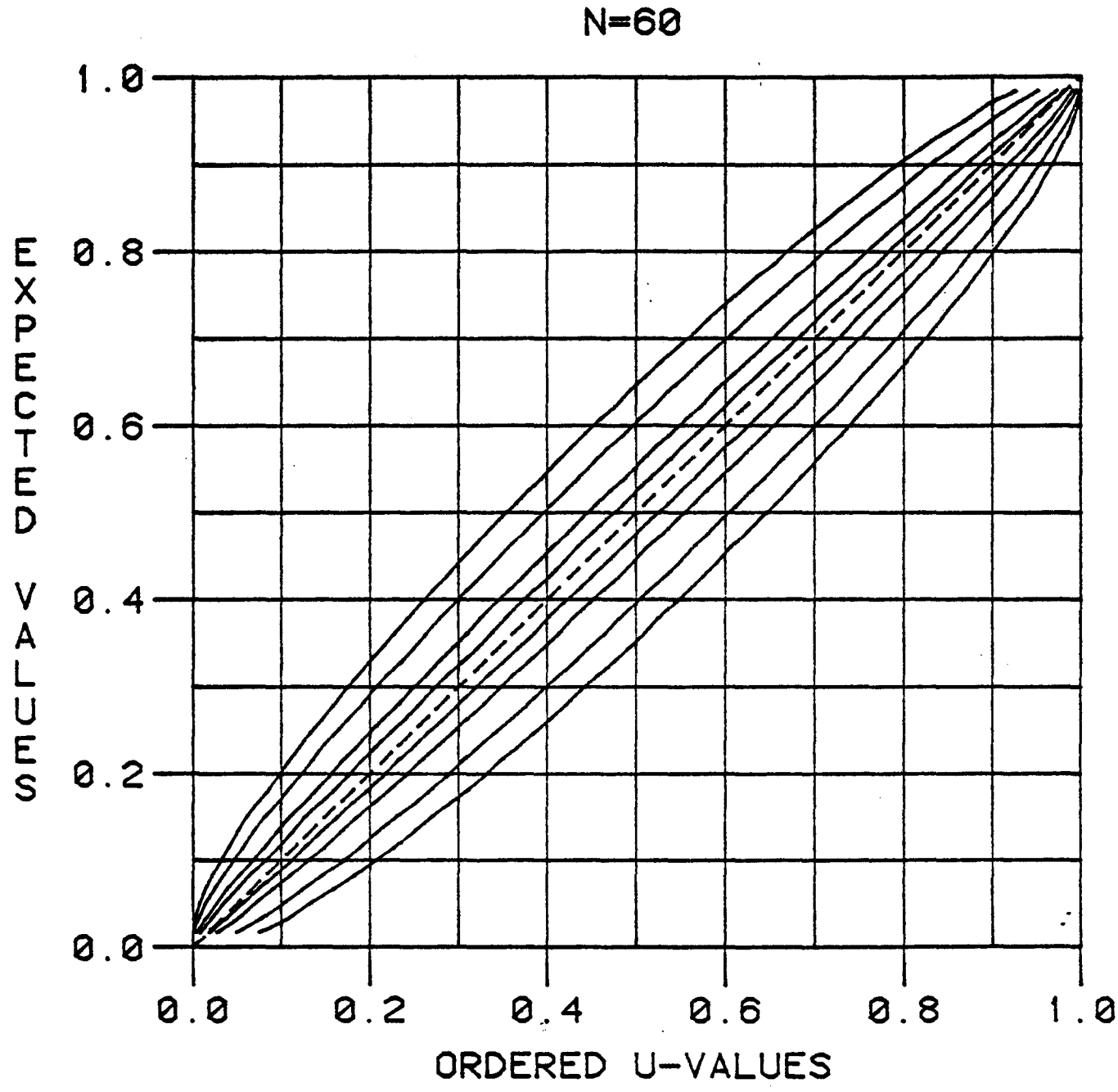

Figure 11: Concentration Bands for N = 60

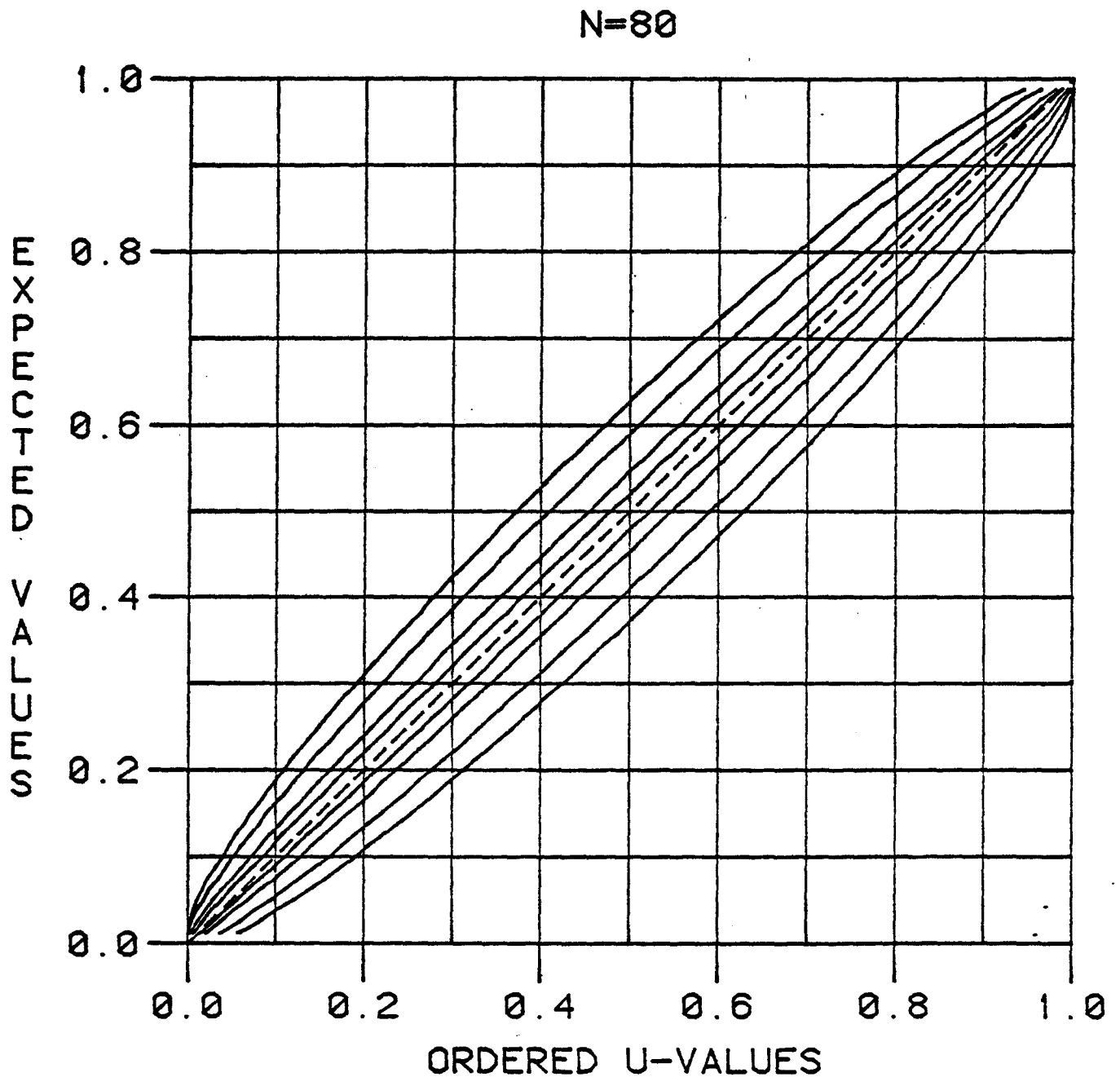

Figure 12: Concentration Bands for N = 80

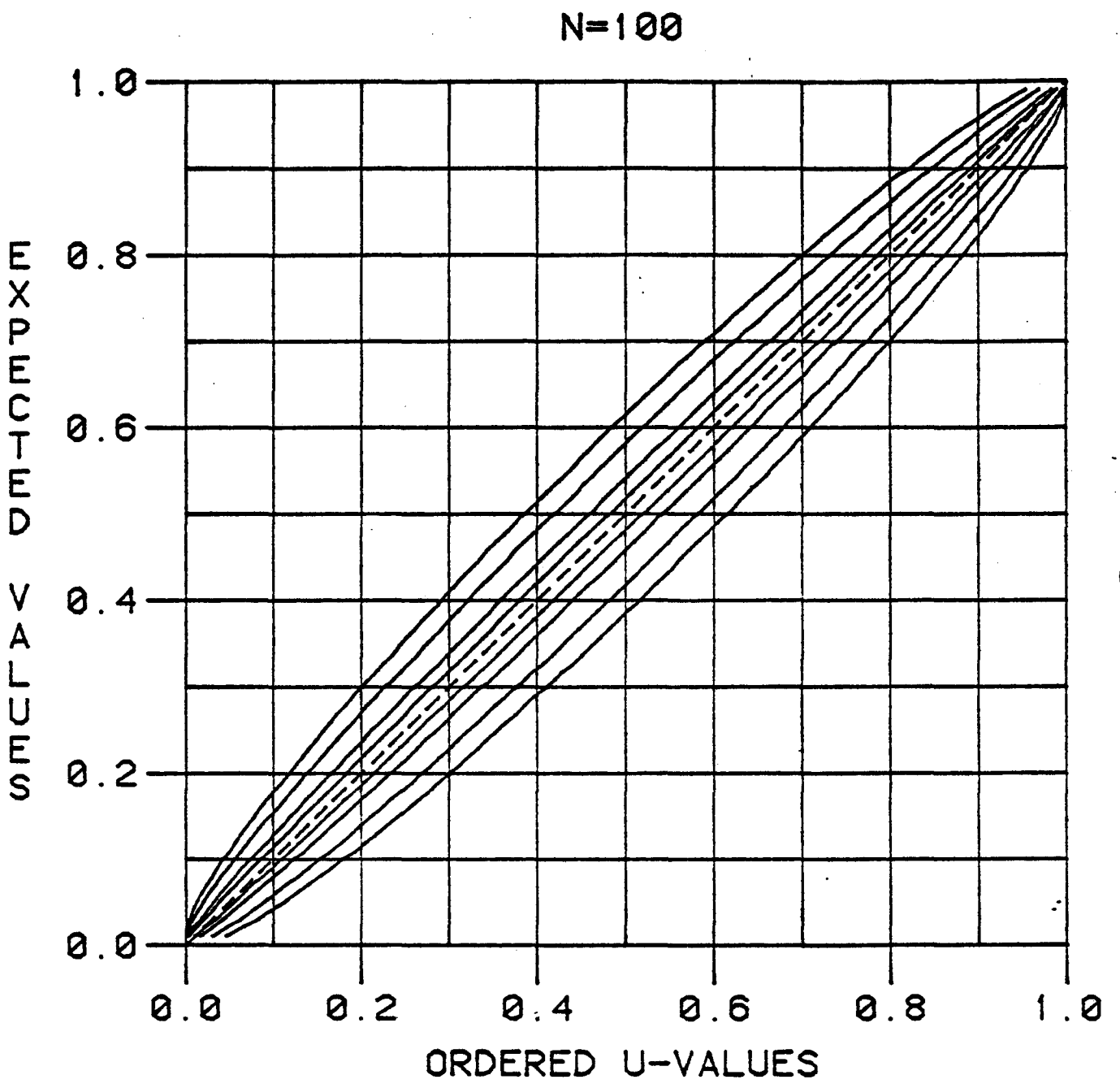

Figure 13: Concentration Bands for N = 100

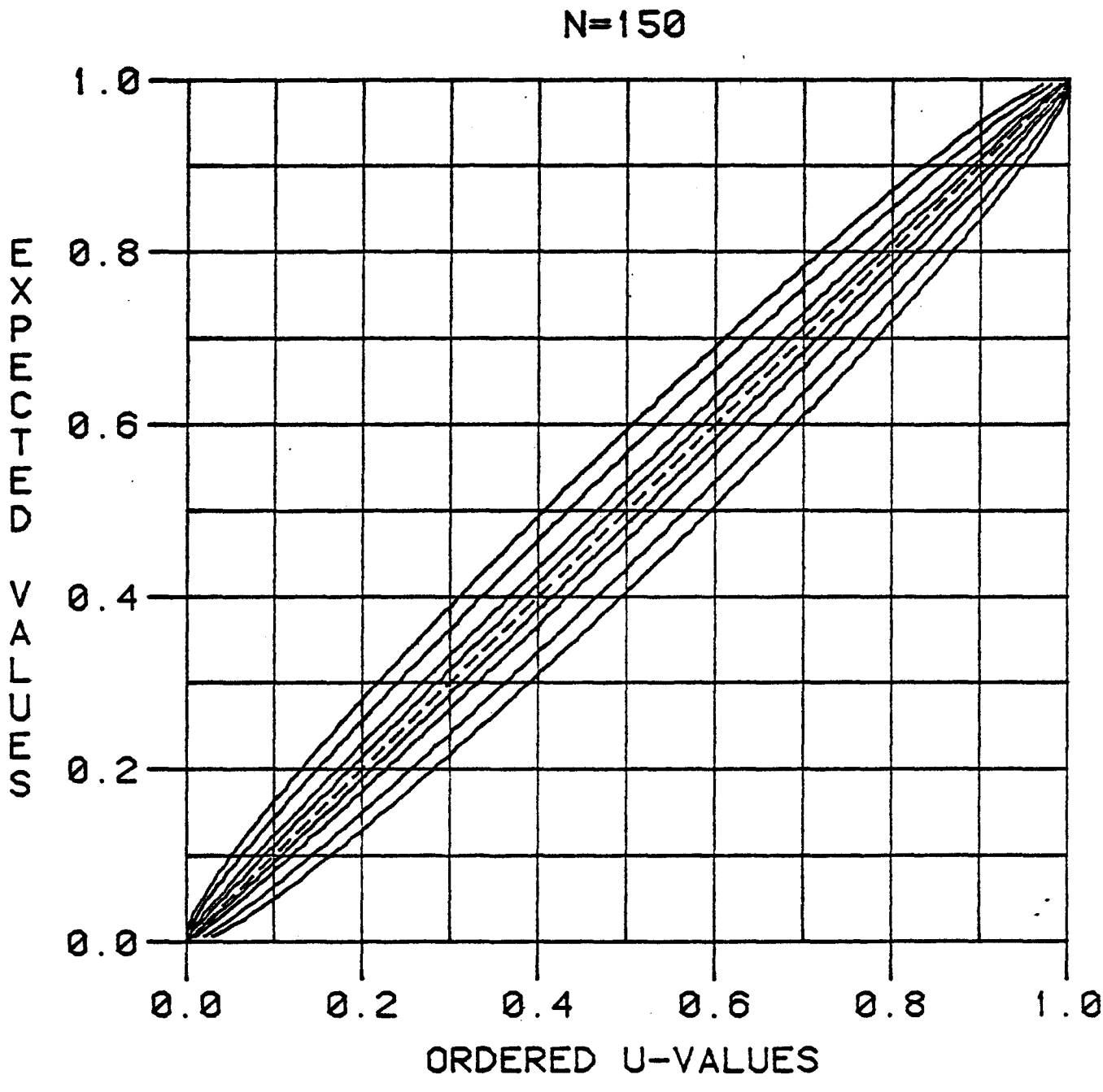

Figure 14: Concentration Bands for N = 150

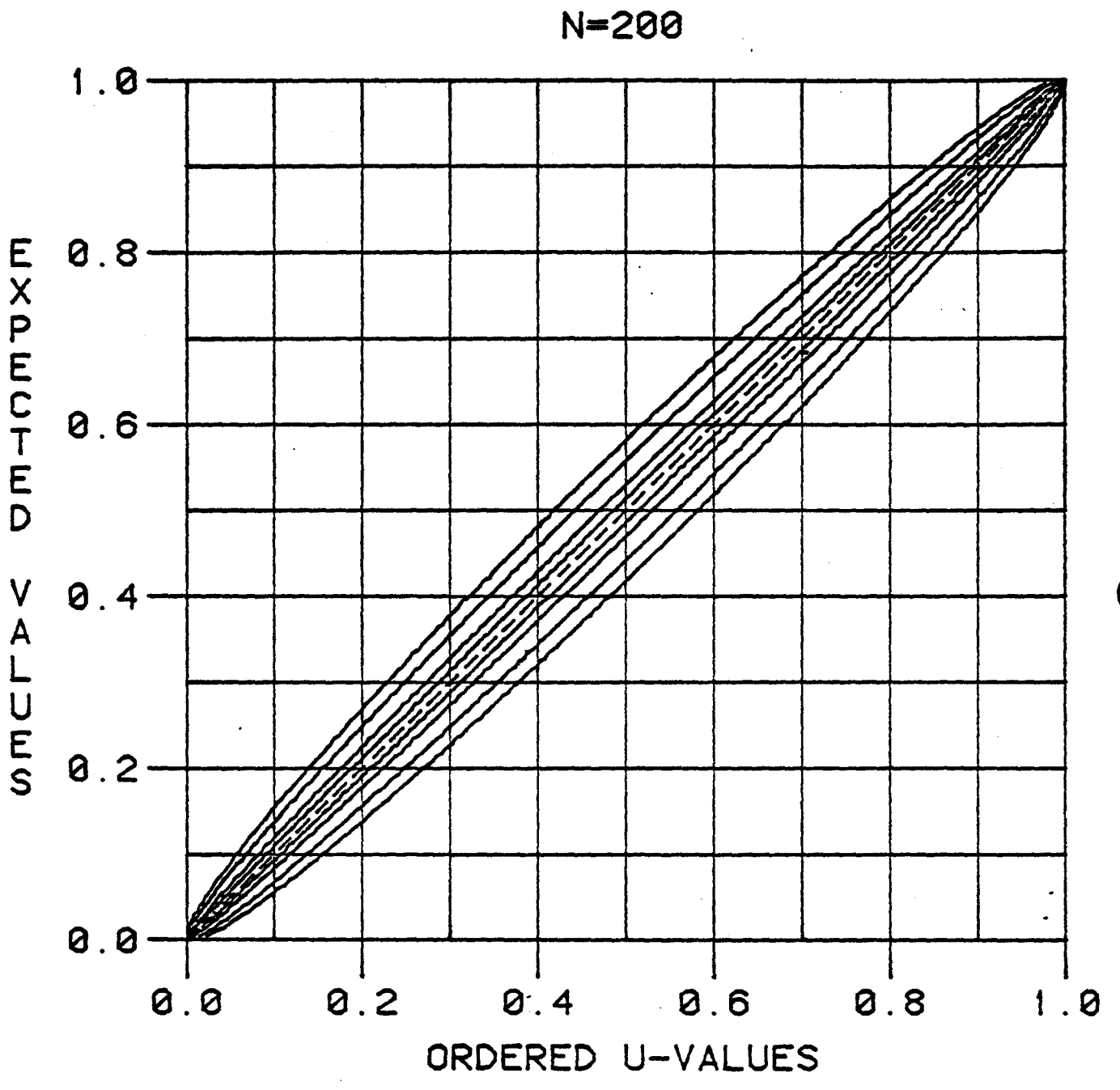

Figure 15: Concentration Bands for N = 200

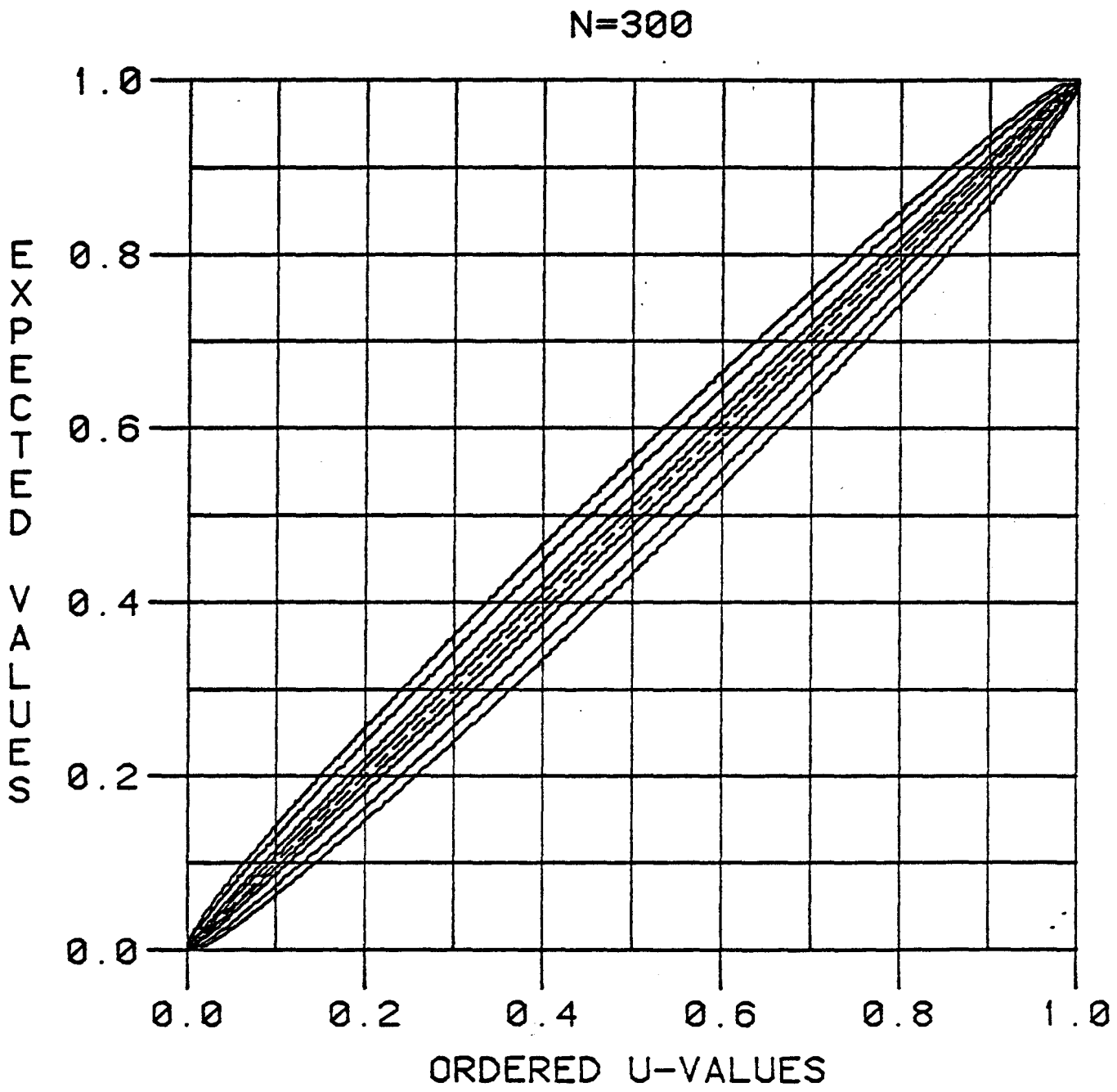

Figure 16: Concentration Bands for N = 300

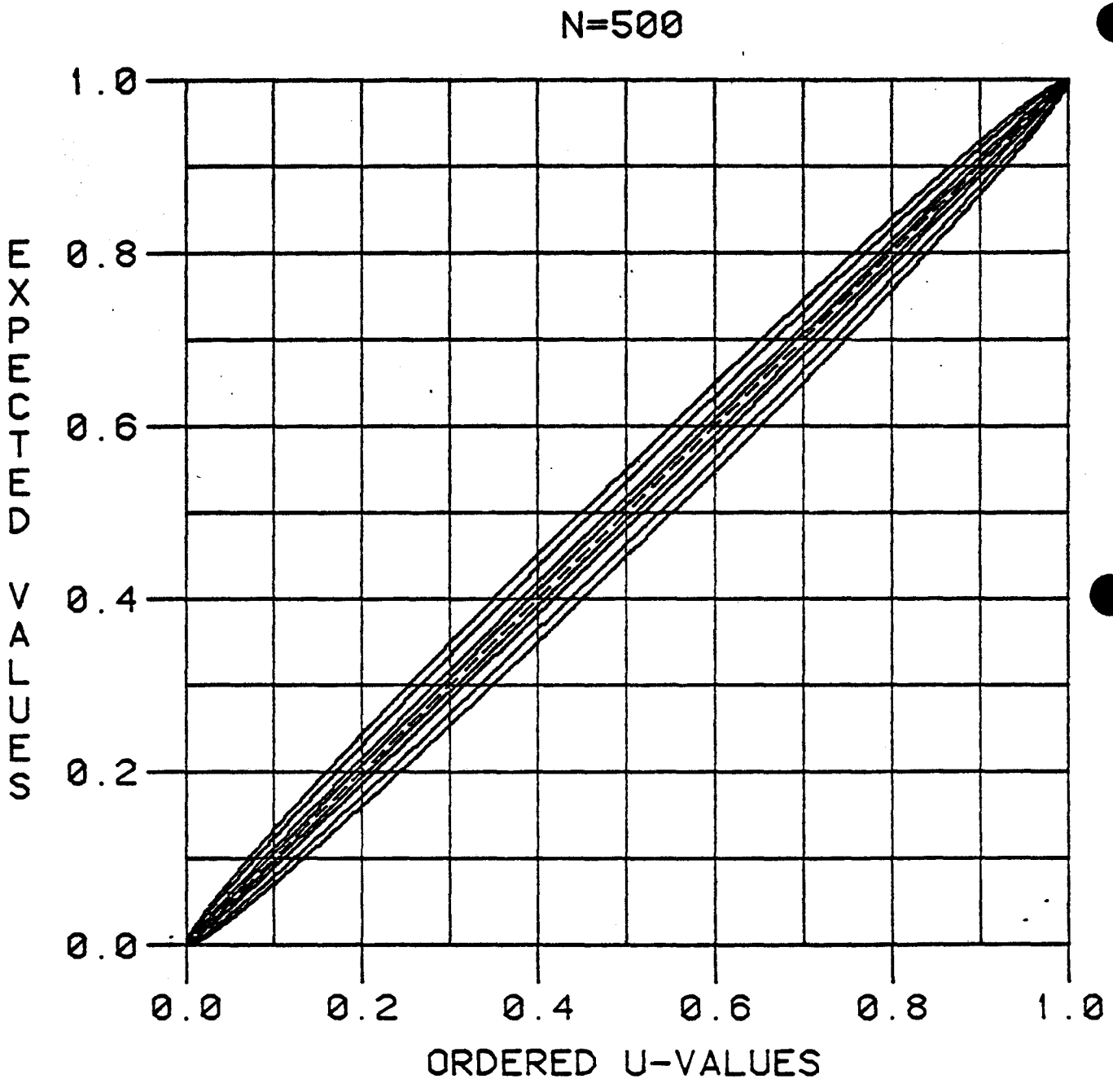

Figure 17: Concentration Bands for N = 500

REFERENCES

Miller, F. L., Jr. and Quesenberry, C. P. (1979). Power Studies of Tests for Uniformity, II. Commun. Statist. - Simula. Computa., B8(3), 271-290.

Neyman, Jerzy (1937). "Smooth" Test for Goodness of Fit. Skandinavisk Aktuarietidskrift, 20, 149-199.

O'Reilly, F. and Quesenberry, C. P. (1973). The conditional probability integral transformation and applications to obtain composite chi-square goodness of fit tests. Ann. Statist., 1, 1-10.

Quesenberry, C. P. (1975). Transforming samples from truncation parameter distributions to uniformity. Commun. Statist., 4(12), 1149-1155.

Quesenberry, C. P. (1979). Some Transformation Methods in Goodness-of-fit, Chap. 6 in Handbook of Goodness-of-fit. Edited by M. A. Stephens and R. D'Agostino. Marcel-Dekker. (to appear)

Quesenberry, C. P. and Miller, F. L., Jr. (1977). Power Studies of Some Tests for Unifromity. J. Statist. Comput. Simul. 5, 169-191.

Quesenberry, C. P. and Starbuck, R. R. (1976). On optimal tests for separate hypotheses and conditional probability integral transformations. Commun. Statist., A5(6), 507-524.

Quesenberry, C. P., Whitaker, T. B., and Dickens, J. W. (1976). On testing normality using several samples: An analysis of peanut aflatoxin data. Biometrics, 32(4), 753-759.

Rincon-Gallardo, S., Quesenberry, C. P. and O'Reilly, F. (1979). Conditional Probability Integral Transformations and Goodness-of-fit Tests for Multivariate Normal Distributions. Ann. of Statist., 7, (to appear).

Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics Without Extensive Tables. J. R. Statist. Soc. B, 32, 115-122.

Watson, G. S. (1961). Goodness-of-fit Tests on a Circle. Biometrika, 48, 109-114.