NOTES

This section is devoted to brief research and expository articles and other short items.

NON-PARAMETRIC TOLERANCE LIMITS[1]

By R. B. Murphy

Princeton University

1. Summary. In this note are presented graphs of minimum probable population coverage by sample blocks determined by the order statistics of a sample from a population with a continuous but unknown cumulative distribution function (c.d.f.). The graphs are constructed for the three tolerance levels .90, .95, and .99. The number, m, of blocks excluded from the tolerance region runs as follows: $m = 1(1)6(2)10(5)30(10)60(20)100$, and the sample size, n, runs from m to 500.

Thus the curves show the solution, β, of the equation $1 - \alpha = I_\beta(n - m + 1, m)$ for $\alpha = .90, .95, .99$ over the range of n and m given above, where $I_x(p, q)$ is Pearson's notation for the incomplete beta function.

Examples are cited below for the one- and two-variate cases. Finally, the exact and approximate formulae used in computations for these graphs are given.

2. Introduction. Suppose a sample of size n is drawn from a population having a continuous cumulative distribution function (c.d.f.), $F(x)$. Let the sample values arranged in order of increasing magnitude be $x_1, x_2, \cdots, x_n$. The fraction, u, of the population which is included between x_r (the r-th smallest value in the sample) and x_{n-s+1} (the s-th largest value) is $F(x_{n-s+1}) - F(x_r)$. This quantity u has been called the *population coverage* for the interval (x_r, x_{n-s+1}). The probability element for this coverage is

$$f(u)\,du = \frac{\Gamma(n+1)}{\Gamma(n-m+1)\Gamma(m)} u^{n-m}(1-u)^{m-1}\,du \tag{2.1}$$

where $m = r + s$. From (2.1) we can calculate the probability that this coverage is at least a given amount, say β. If we call this probability α, we have

$$\alpha = \int_\beta^1 f(u)\,du. \tag{2.2}$$

The quantity α is the probability that $100\beta\%$ of the population will be included between x_r and x_{n-s+1}, and it is called the *tolerance level*. This probability depends only on n and m $(=r + s)$.

[1] All computations involved in this paper were carried out under an Office of Naval Research contract.

The idea of coverage is more general than it first appears. If we think of $x_1, x_2, \cdots, x_n$ as points plotted along the x-axis, we will then have $n + 1$ intervals: $(-\infty, x_1), (x_1, x_2), \cdots, (x_n, +\infty)$, which, following Tukey [3], we will call *blocks*. The reason for this term will be clear when we deal with the case of a sample from a population of more than one variable. The coverage for the i-th block (x_i, x_{i+1}) is $F(x_{i+1}) - F(x_i)$. The probability element of the sum of the coverages of *any* preassigned group of $n - m + 1$ blocks is given by (2.1) and hence the probability α that the fraction of the population covered by *any* $n - m + 1$ blocks is given by (2.2). By preassigned blocks we mean ones designated by order statistics prior to obtaining any sample from which a prediction is to be made with these blocks. In general it is *not* legitimate, after taking a sample and for some reason evident only then, to specify which blocks in this sample are to be included or excluded from the coverage. There is no objection, however, to specifying a scheme of blocks for the coverage on the basis of past samples when the scheme is to be applied to future samples.

The purpose of this note is to present graphs of β as a function of n for $m = 1(1)6(2)10(5)30(10)60(20)100$ and for $\alpha = .90, .95, .99$. There are three figures: Figure 1 gives curves for $\alpha = .90$, Figure 2 for $\alpha = .95$, and Figure 3 for $\alpha = .99$. The graphs are accurate to at least two decimal places but never more than three. In terms of the Pearson notation (2.2) gives, after minor alternation, $1 - \alpha = I_\beta\ (n - m + 1, m)$. Hence these graphs may also be used to find the 10, 5 and 1 per cent points of a variate X $(0 \leq X \leq 1)$ with the c.d.f. $I_x(p, q)$ for $1 \leq p \leq 500$ and $1 \leq q \leq 100$.

3. Computations for the graphs.

If in the relation (2.2) three of the arguments α, β, m, and n are given, the solution for the fourth may often be found in Pearson [5] or Thompson [6]. The values of β through $n = 100$ were computed exactly for these graphs. For larger n, β was computed approximately from

$$\beta \cong \left[\frac{\sqrt{(\chi^2_\alpha - 2m)^2 + 16n(n - m)} - (\chi^2_\alpha - 2m)}{4n}\right]^2 \tag{3.1}$$

where χ^2_α is determined by the relation

$$Pr(\chi^2 \geqq \chi^2_\alpha) = 1 - \alpha$$

and has $2m$ degrees of freedom. This approximation is due to Scheffé and Tukey. For large m the Cornish-Fisher approximation to χ^2_α was used.

4. Illustrations of the one-variate case.

The most common use to which the graphs presented here may be put is in the prediction of β in sampling from a distribution of a single random variable. It is this case that was first presented by Wilks [1]. Suppose in the mass production of a certain type of screw one is interested in the least proportion of all screws manufactured that have lengths between the least and greatest lengths appearing in a random sample of 100

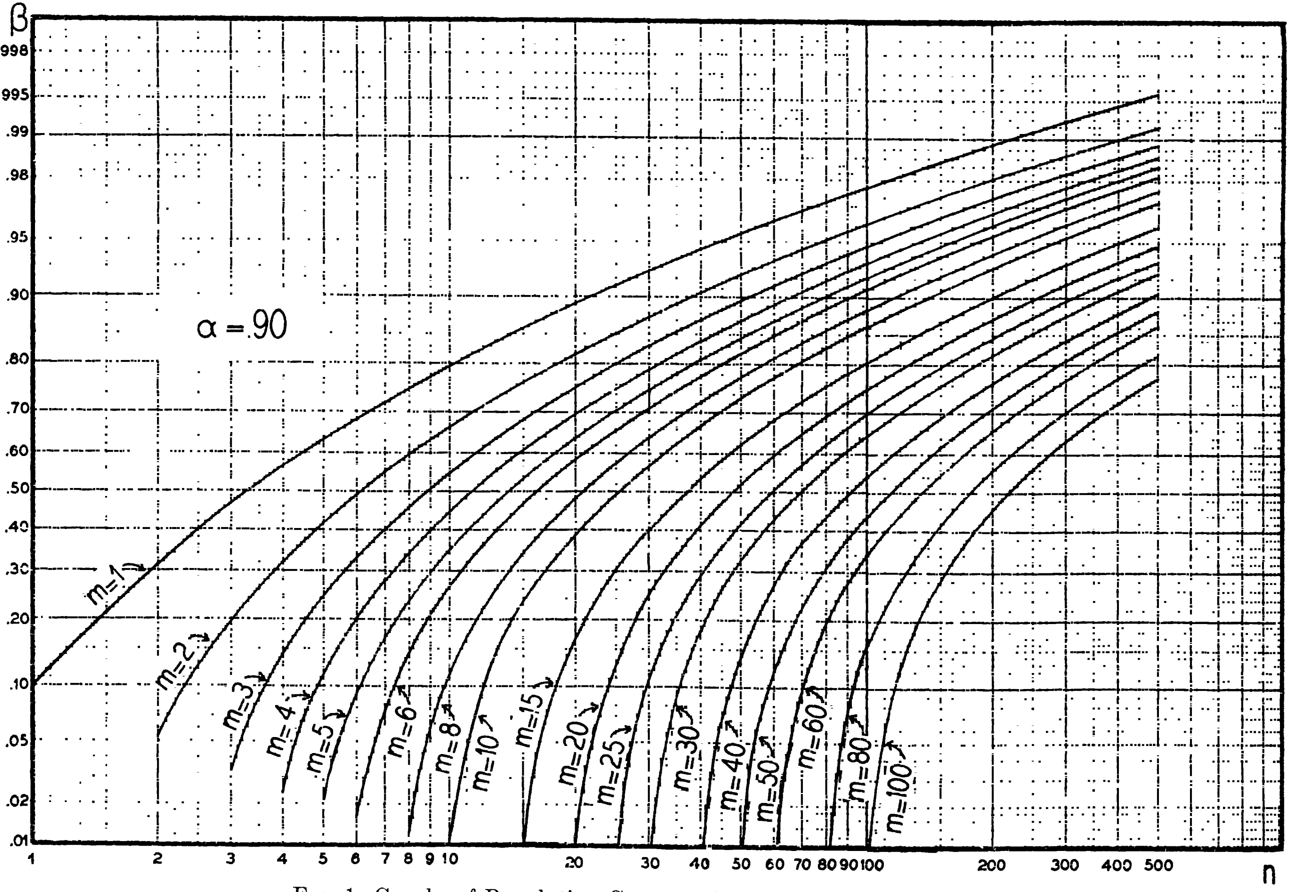

FIG. 1. Graphs of Population Coverage for the Tolerance Level .90.

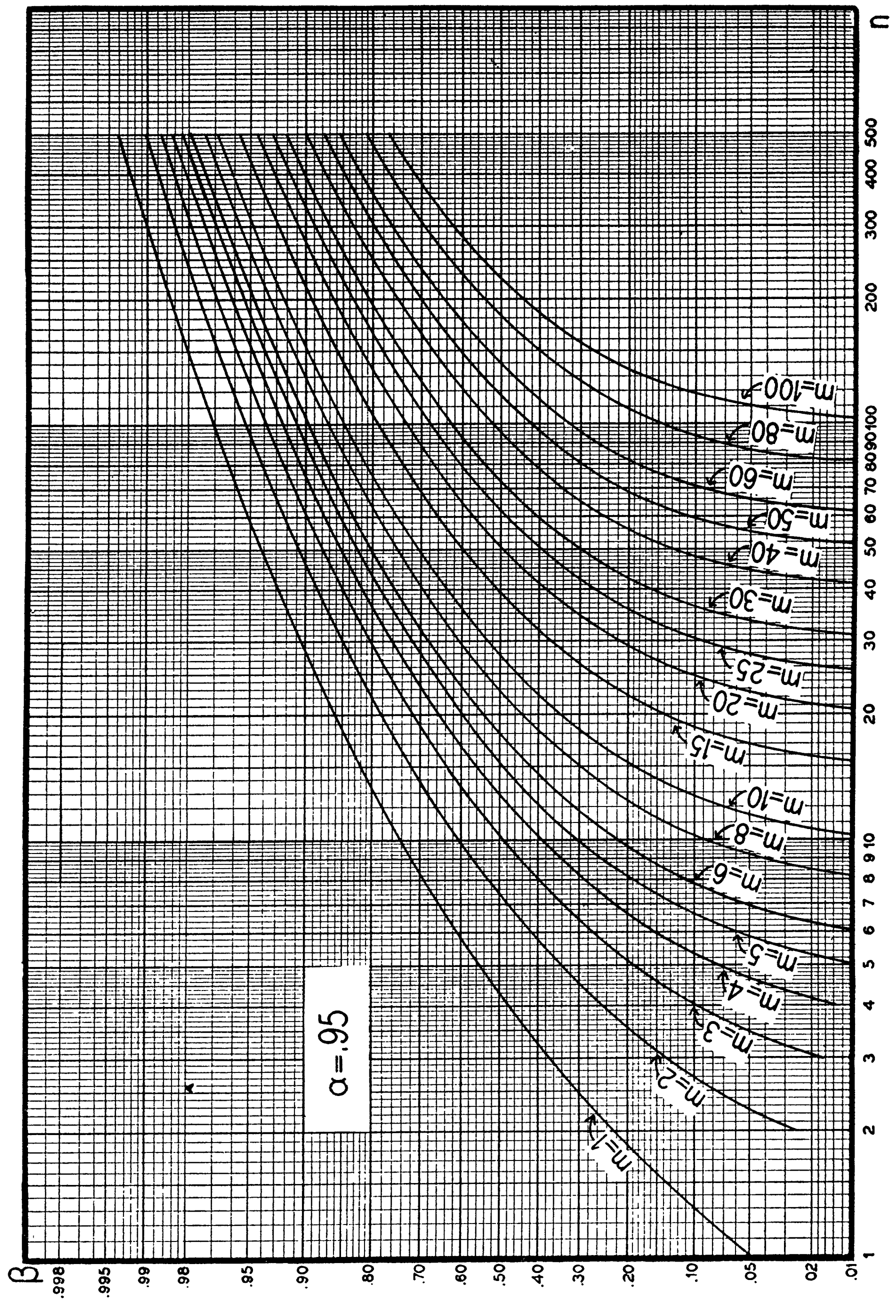

FIG. 2. Graphs of Population Coverage for the Tolerance Level .95.

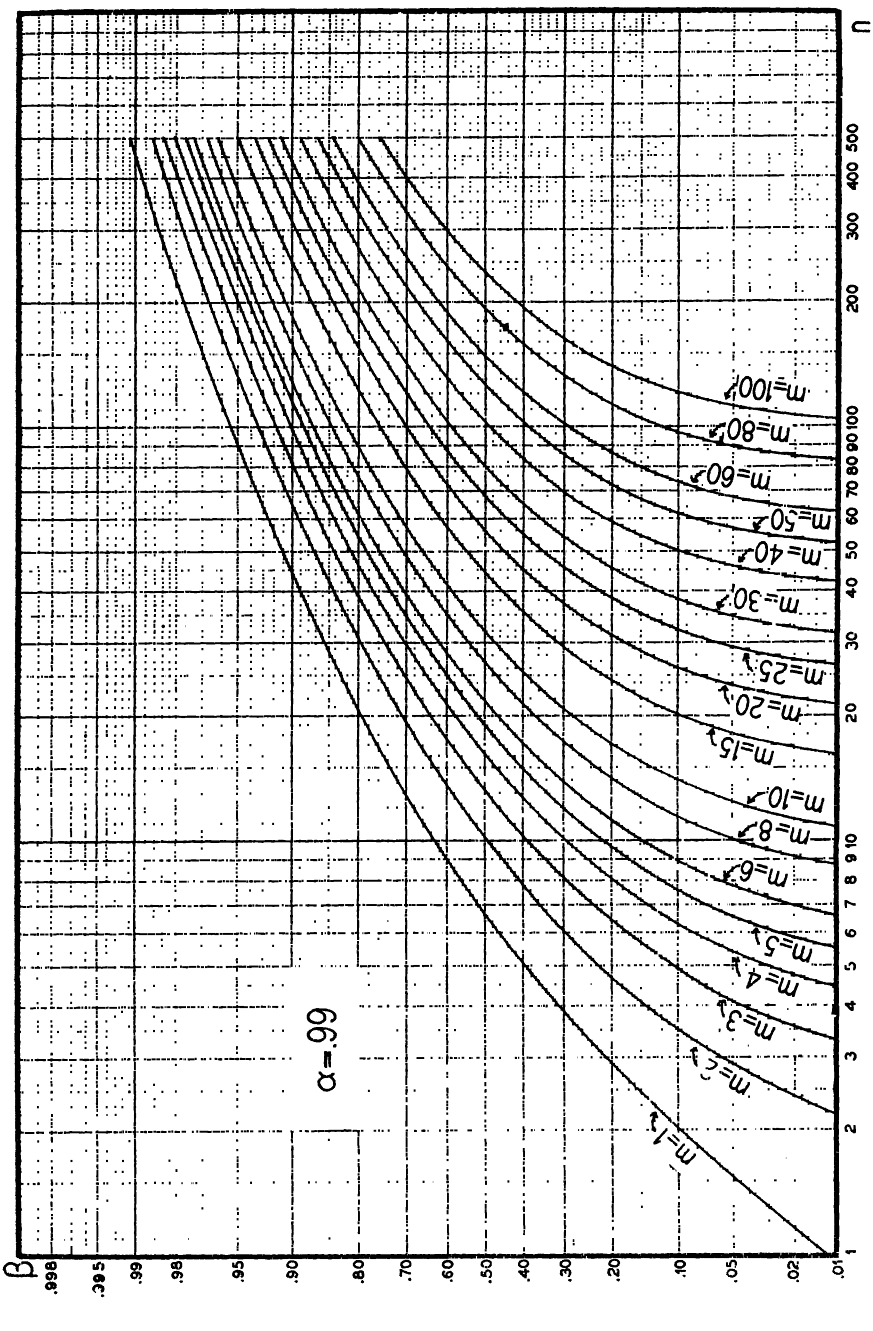

FIG. 3. Graphs of Population Cov :rage for the Tolerance Level .99.

screws. It is assumed that we do not know the distribution of the length, X, of a screw produced in this process. Furthermore, it is assumed, of course, that the manufacturing process is in a state of statistical control in the sense of Shewhart. We plan to discard two blocks: $(-\infty, x_1)$ and $(x_{100}, +\infty)$—exactly as many blocks as observations. At the level $\alpha = .99$ we obtain from Figure 3 that at least 93.5% of all screws in the population sampled have lengths that fall between x_1 and x_{100}. If we now draw a random sample of 100 screws and find the least and greatest screw lengths to be 1.40 and 1.60 inches respectively, we may say that at least 93.5% of all screws from the population sampled have lengths between 1.40 and 1.60 inches at the .99 tolerance level. It must be observed that the prediction is made on the basis of preassigned order statistics, and not of the values 1.40 and 1.60.

We might equally as well have put the question in another way: If we want at least 93.5% of the lengths of all screws to lie within the range of lengths of a sample of 100 screws, then at the tolerance level $\alpha = .99$ what is the smallest sample we could have in which as many as 2% of the sample are not acceptable? Examining the intersections of the curves in Figure 3 with the line $\beta = .935$ we choose the smallest n such that $m/n \leqq .02$ and find $n = 100$.

5. The case of more than one variate. The ideas given in the introduction may be extended to sampling situations involving two or more statistically dependent variates with a continuous joint c.d.f. by means of the notion of blocks. The abstract formulation is given by Tukey [3]. We shall restrict ourselves to the case of two dependent variates X and Y, but the generalization is obvious. Because of the dependence, the joint population of X and Y may be expressed as an associated pair of values $W = (X, Y)$. Suppose a sample of size n is drawn from this population, and let the pairs be $w_1, w_2, \cdots, w_n$, where $w_i = (x_i, y_i)$. If we now choose a sequence of n numerically valued functions of x and y (or of w), $f_1(w), \cdots, f_n(w)$, let us order the w_i in a sequence $w_1^{(1)}, w_2^{(1)}, \cdots, w_n^{(1)}$ such that $f_1(w_{i+1}^{(1)}) > f_1(w_i^{(1)})$. Imagine now that the sample values are plotted in a plane scatter diagram. We call the first block the set of points $w = (x, y)$ such that $f_1(w) < f_1(w_1^{(1)})$. That is, we may imagine the curve $f_1(x, y) - f_1(w_1^{(1)}) = 0$ plotted in the plane and that the first block is bounded by this curve. Then discarding $w_1^{(1)}$ we take the $n - 1$ remaining w_i and order them in a sequence $w_1^{(2)}, w_2^{(2)}, \cdots, w_{n-1}^{(2)}$ such that $f_2(w_{i+1}^{(2)}) > f_2(w_i^{(2)})$. We call the second block the set of points $w = (x, y)$ such that $f_1(w) \geq f_1(w_1^{(1)})$ and also $f_2(w) < f_2(w_1^{(2)})$. Thus the second block is bounded by the curves $f_1(x, y) - f_1(w_1^{(1)}) = 0$ and $f_2(x, y) - f_2(w_1^{(2)}) = 0$. If we continue this process of discarding and reordering, until all n functions f_i are used, we shall obtain a division of the plane into $n + 1$ non-overlapping blocks, the "extra" block arising at the last step in the process. Then the fraction, u, of "points" (X, Y) of the joint population of X and Y that are covered by *any* $n - m + 1$ blocks has the probability element (2.1). Also the probability α that the population coverage, u, will be at least as large as β is given by (2.2). The $n - m + 1$ blocks constitute a tolerance *region*.

An extension of this case has been made by Wald [2]. Namely, before a sample is taken let us choose a numerically valued function f of w and choose $k(\leqq n)$ of the w_i and order them in a sequence $w_{a_1}^{(0)}, w_{a_2}^{(0)}, \cdots, w_{a_k}^{(0)}$ such that $f(w_{a_{j+1}}^{(0)}) > f(w_{a_j}^{(0)})$ and $a_{j+1} > a_j$. Next, within each "strip" of the (x, y) plane such that $w = (x, y)$ satisfies $f(w_{a_{j+1}}^{(0)}) > f(w) > f(w_{a_j}^{(0)})$, suppose that we follow the construction in the previous paragraph. Then the population coverage, u, by $n - m + 1$ blocks from one or more of these strips or their exteriors has the probability element (2.1).

Again the warning must be made that the above functions $f, f_1, f_2, \cdots, f_n$, the numbers $a_1, a_2, \cdots, a_k$ and the sequence of construction must be completely specified before samples are drawn to which this scheme is to be applied.

6. Illustrations for two variates. As an example of the use of the graphs for a two-variate case, we use an example cited by Tippett [8]. The two variates are the percentage of pig iron, X, and the lime consumption, Y, per cwt. of steel in 100 steel castings made without slag control. A scatter diagram is given in Figure 4. Unfortunately the value of this example is lessened by the fact that the block schemes were made after the sample had been taken; it does illustrate, at least, the two simple types of scheme.

The tolerance region T (solid lines in Figure 4) resulted from the following scheme: let $f_1(w) = y, f_2(w) = f_3(w) = f_4(w) = f_5(w) = f_6(w) = -y$. Now follow the Wald procedure choosing $f(w) = y$ with $k = 6$, and $a_1 = 1, a_2 = 13, a_3 = 46, a_4 = 75, a_5 = 90, a_6 = 96$. Then in each strip $y_{a_{j+1}} > y > y_{a_j}$ let $f_\iota(w) = x$. Considering only the blocks within the heavy line as the tolerance region, we have, by counting the discarded blocks, $m = 16$.

In constructing the region T' (broken lines in Figure 4) we also use Wald's method, taking $f(w) = y - 5x$ with $k = 2$ and $a_1 = 3, a_2 = 96$. In the exterior region with $f(w) > f(w_{96}^{(0)})$ let all $f_i = y + 5x$ and similarly in the exterior region $f(w) < f(w_3^{(0)})$. Then in the strip $f(w_{96}^{(0)}) > f(w) > f(w_3^{(0)})$ (i.e., in the region in which $41 > y - 5x > -77$) choose $f_1(w) = y, f_2(w) = f_3(w) = f_4(w) = -y, f_5(w) = f_6(w) = f_7(w) = y + 5x$, and $f_8(w) = f_9(w) = -y - 5x$. Counting the blocks outside the heavily bordered region, we have $m = 17$.

We obtain by interpolation $\beta = .80$ for T and $\beta = .78$ for T' at the $\alpha = .90$ level.

7. Ties. A tie is a sample point which in a coordinate system defining a set of order statistics coincides in one or more coordinates with other sample points. For instance, in the X coordinate of our example (32, 159) and (32, 185) are tied, and (47, 218) and (47, 218) are tied in any system of coordinates. It would seem easier to avoid ties with regions of the type of T' than with those of the type of T.

The existence of ties in the population is assumed impossible, because positive point probabilities would destroy the continuity of the c.d.f. Therefore we attribute the ties to the crudity of measuring devices.

A procedure for handling ties is given by Tukey [4].

8. Acknowledgments. The author wishes gratefully to acknowledge the assistance of Dr. S. S. Wilks in the preparation of this note and of Dr. J. W. Tukey, who also suggested the data used in section 6.

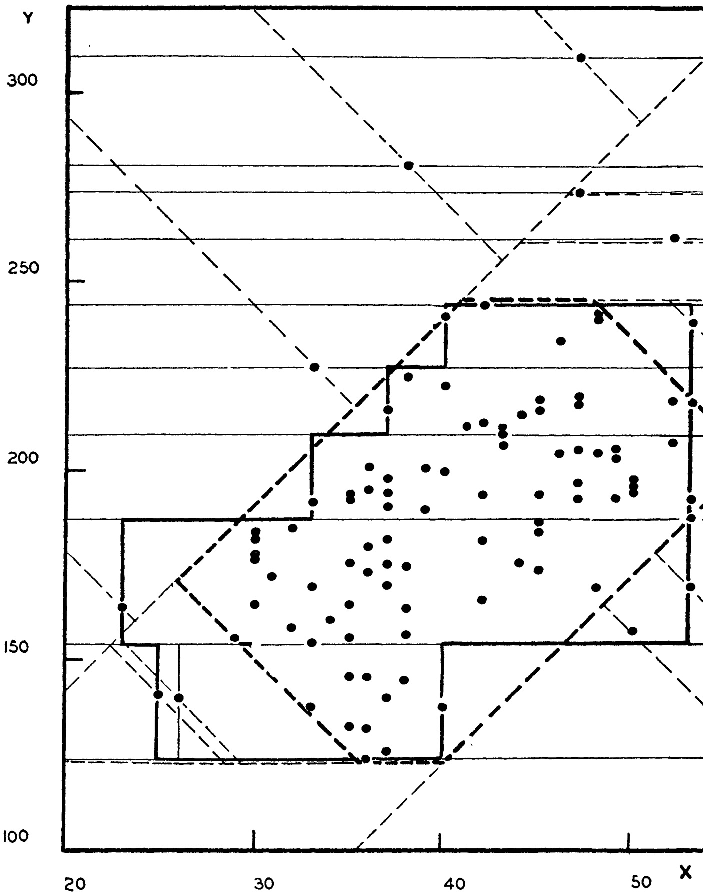

FIG. 4. Illustrative Tolerance Regions for Two Variates.

REFERENCES

[1] S. S. Wilks, "Statistical prediction with special reference to the problem of tolerance limits," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 400–409.

[2] A. Wald, "An extension of Wilks' method for setting tolerance limits," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 45–55.

[3] J. W. Tukey, "Non-parametric estimation II. Statistically equivalent blocks and tolerance regions—the continuous case," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 529–539.

[4] J. W. Tukey, "Non-parametric estimation III. Statistically equivalent blocks and multivariate tolerance regions—the discontinuous case," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 30–39.
[5] K. Pearson, *Tables of the Incomplete Beta-Function*, Cambridge, 1934.
[6] C. M. Thompson, "Tables of percentage points of the incomplete beta function," *Biometrika*, Vol. 32, Part II (1941), pp. 151–181.
[7] H. Goldberg and H. Levine, "Approximation formulas for the percentage points and normalization of t and χ^2", *Annals of Math. Stat.*, Vol. 17 (1946), pp. 216–225.
[8] L. H. C. Tippett, *Statistical Methods in Industry*, Iron and Steel Industrial Research Council, British Iron and Steel Federation, 1943.

THE FOURTH DEGREE EXPONENTIAL DISTRIBUTION FUNCTION[1]

By Leo A. Aroian

Hunter College

We shall derive a recursion formula for the moments of the fourth degree exponential distribution function, state its more characteristic features, and show how the graduation of observed distributions may be accomplished by the method of moments and the method of maximum likelihood. The purpose of the note is to make possible a wider use of this function.

R. A. Fisher [1] introduced the fourth degree exponential function

$$y_t = k \exp\{-(\beta_4 t^4 + \beta_3 t^3 + \beta_2 t^2 + \beta_1 t)\}, \tag{1}$$

where $r_1 \leq t \leq r_2$, $t = (x - m)/\sigma$, m indicates the population mean, σ the population standard deviation, and where the β's are functions of

$$\alpha_n = \int_{r_1}^{r_2} t^n y_t \, dt.$$

A. L. O'Toole in two stimulating papers [2], [3], has studied (1); however his methods and results are unnecessarily complicated. O'Toole requires eight moments to determine parameters similar to the β's. Both Fisher and O'Toole considered the restricted class of (1) with range $(-\infty, \infty)$.

Let

$$u = t^n \exp\{-(\beta_4 t^4 + \beta_3 t^3 + \beta_2 t^2)\}, \; dv = e^{-\beta_1 t} \, dt \tag{2}$$

in

$$\alpha_n = \int_{r_1}^{r_2} t^n y_t \, dt, \quad \text{obtaining} \tag{3}$$

$$4\beta_4 \alpha_{n+3} + 3\beta_3 \alpha_{n+2} + 2\beta_2 \alpha_{n+1} + \beta_1 \alpha_n = n\alpha_{n-1}, \qquad n = 1, 2, 3, \cdots, \tag{4}$$

[1] Presented to the American Mathematical Society and the Institute of Mathematical Statistics, September 4, 1947.

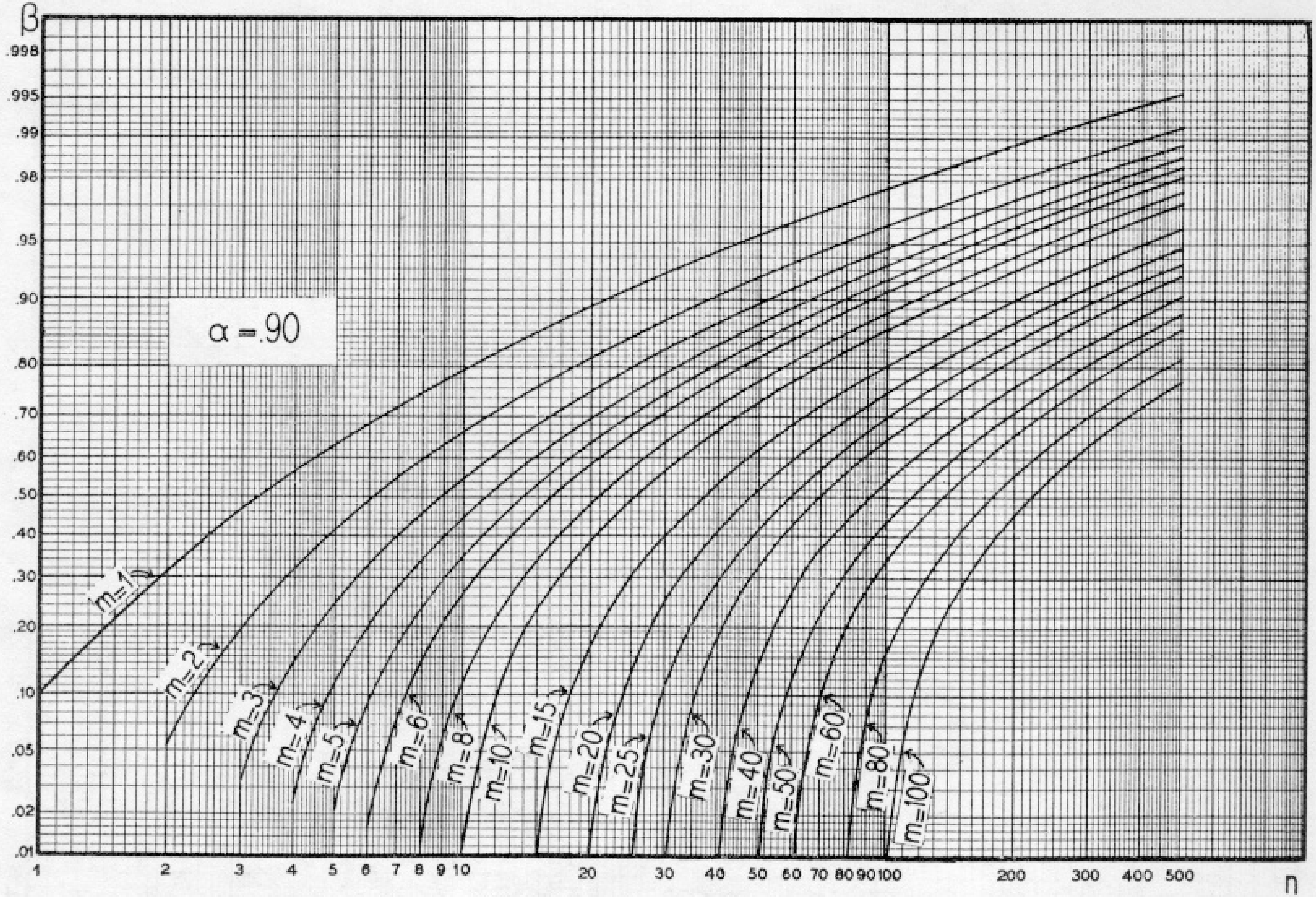

FIG. 1. Graphs of Population Coverage for the Tolerance Level .90.

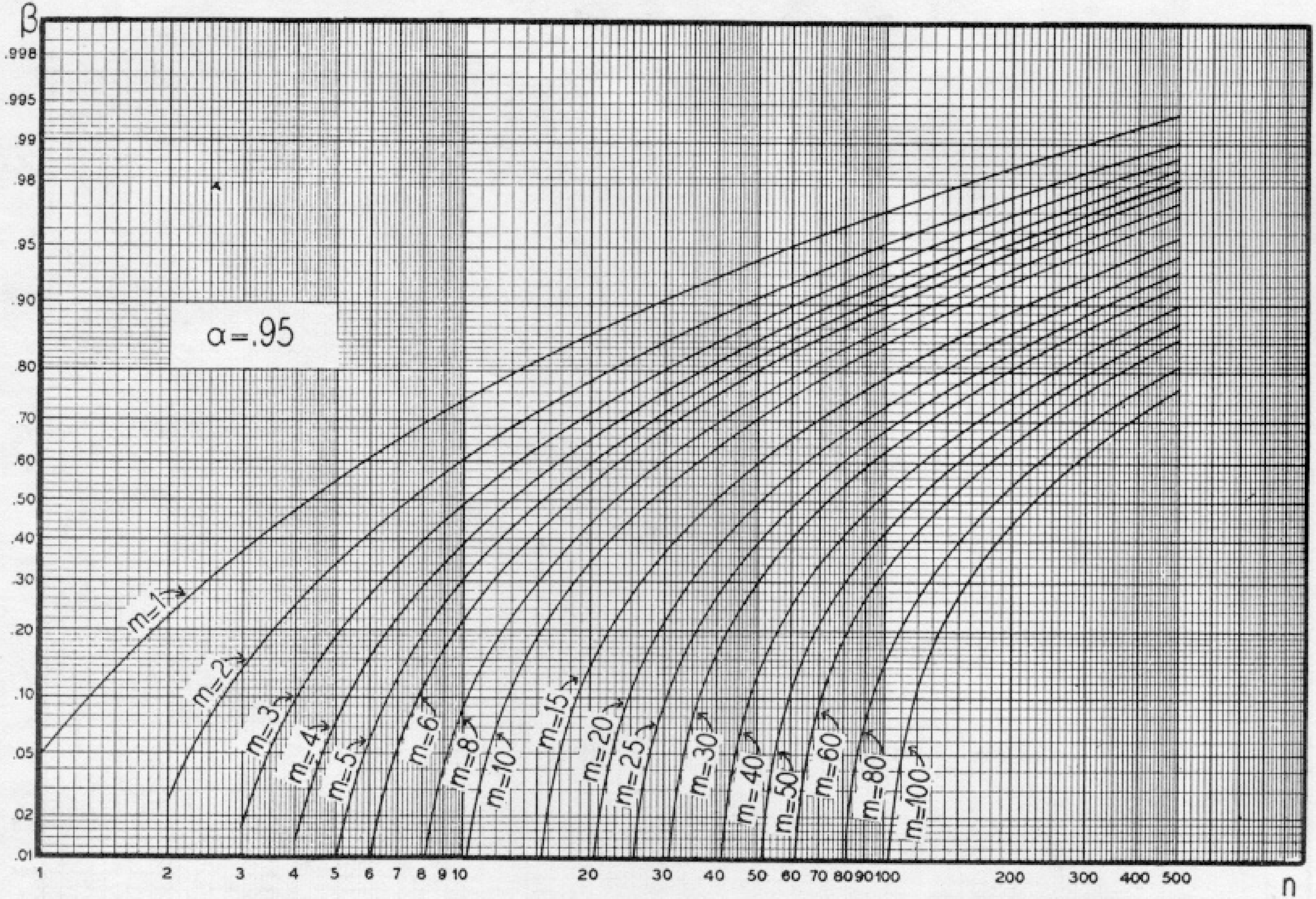

FIG. 2. Graphs of Population Coverage for the Tolerance Level .95.

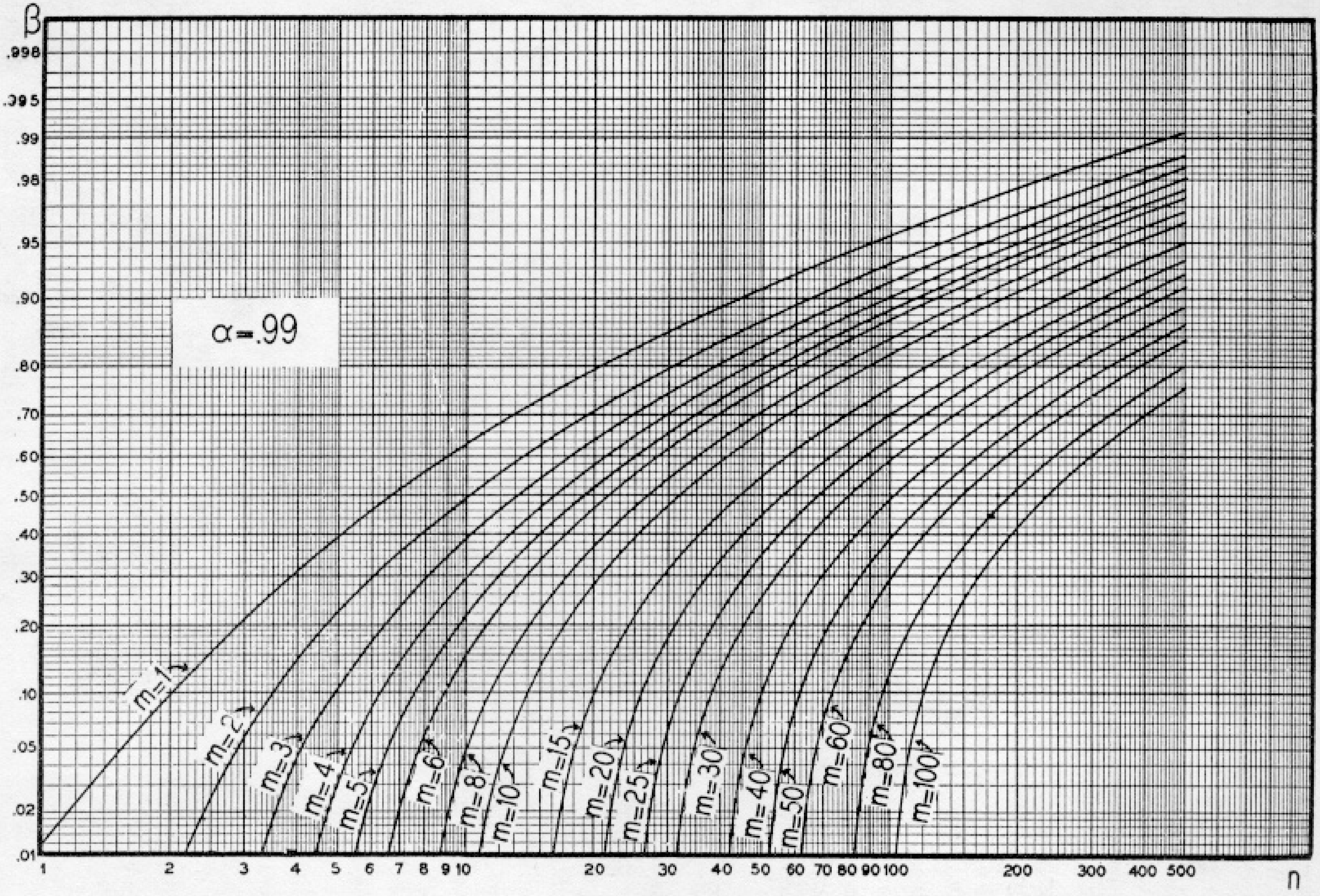

FIG. 3. Graphs of Population Coverage for the Tolerance Level .99.