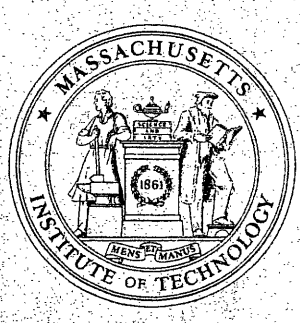
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

THE GEOMETRIC AND THE BRADFORD DISTRIBUTIONS, A COMPARISON

by

Philip M. Morse

OR 049-76 February 1976

Supported in part by the U.S. Army Research Office (Durham) under Contract No. DAHC04-73-C-0032.

The Geometric and the Bradford Distributions, a Comparison.

by Philip M. Morse
Operations Research Center,
Mass. Inst. of Technology

Abstract

Both the geometric and the Bradford probability distributions are used to describe collections of items of interest in information science. Each unit item has a productivity, an integer n measuring the amount of use of the item. The cumulative fraction F_n of items with productivity equal to n or greater may be expressed as a function of n or else as a function of the cumulative mean productivity G_n of items with productivity equal to n or greater. If F_n is an exponential function of n, the distribution is geometric; if it is an exponential function of G_n, it is a Bradford distribution. The exact solution of F_n as a function of n for the Bradford distribution is computed; the results are tabulated. Graphs are given, comparing the two distributions, and their relative usefulness is discussed.

Definitions

Two probabilistic distribution functions are in common use as models to describe library and informational data; the geometric distribution[1] and the Bradford distribution[2,3]. Both have their uses in compressing a large amount of data in terms of a few numerical parameters, from which one can calculate a number of general properties of the collection from which the data were taken. As with any probabilistic situation, neither of these models is expected to fit the data precisely, but both of them have been found to fit a number of cases well enough to make them useful tools for analysis and planning. This paper compares the two distributions in detail, to see how they are related and to point out where each is useful.

Both distributions can model a variety of collections; a collection of books in a library; a collection of journals; or a list of journal articles in a specified field, for example. Each of the items in a given collection has a certain measureable productivity n, an integer, that varies from item to item. A book has circulated n times in the past Q years, for example, or one of the journals has published n articles on some given subject in the past N years, or a journal article in the list has been referenced in n other articles during Q years after its publication. In any of these examples, and in many otners, one can find, by counting, the fraction f_n, of all the items in the collection, that have productivity n and, by further addition, the fraction F_n that have productivity equal to or greater than n.

For example, one can imagine arranging the items in order of productivity, leaving out the completely non-productive items; starting with the fraction f_1 of the productive items that have unit productivity, then the fraction f_2 that have productivity 2 and so on. If there are N_p productive items in all, then $N_p f_1$ items have unit productivity and account for $N_p f_1$ production, $N_p f_2$ items have productivity 2 and account for $2N_p f_2$ production and so on. Since the f_n's are fractions of the total number of productive items, the sum of the f_n's over n, starting with $n = 1$, is unity, so that the sum of all the productive items is N_p,

$$\sum_{n=1}^{\infty} N_p f_n = N_p \qquad \text{since} \qquad \sum_{n=1}^{\infty} f_n = 1 \tag{1}$$

The total production, the sum of $N_p f_1$ plus $2N_p f_2$, etc., is then

$$\sum_{n=1}^{\infty} n N_p f_n = \bar{n} N_p \qquad \text{where} \qquad \bar{n} = \sum_{n=1}^{\infty} n f_n \tag{2}$$

$\bar{n} N_p$ being the total production of all the productive items and $\bar{n}$ being the average productivity per item.

The fraction f_n, as function of n, is the distribution function. The cumulative distribution

$$F_n = \sum_{m=n}^{\infty} f_m \quad ; \quad F_1 = 1 \quad ; \quad F_n \xrightarrow[n\to\infty]{} 0 \tag{3}$$

is the fraction of items with productivity equal to n or greater. The average productivity of this fraction is

$$P_n = G_n / F_n \quad \text{where} \quad G_n = \sum_{m=n}^{\infty} m f_m \tag{4}$$

$$G_1 = \bar{n} \quad ; \quad G_n \xrightarrow[n\to\infty]{} 0$$

The quantity $x_n = G_n/\bar{n}$ is thus the fraction of the total production carried by those items with productivity equal to or greater than n.

The Geometric Distribution.

In a number of cases[1], the data approximately correspond to a geometrical (or exponential) dependence of F_n on n,

$$F_n = \gamma^{n-1} = \exp\left[(n-1)\ln\gamma\right] \quad \text{or} \quad \ln F_n = (n-1)\ln\gamma$$

$$f_n = F_n - F_{n+1} = (1-\gamma)\gamma^{n-1} \; ; \quad G_n = (1-\gamma)\sum_{m=n}^{\infty} m\gamma^{m-1} \qquad (5)$$

$$G_1 = \bar{n} = \frac{1}{1-\gamma} \; ; \quad n = 1, 2, 3, \cdots$$

Since the logarithm of F_n depends linearly on n, the successive values of F_n, plotted on semi-logarithmic paper, lie on a straight line, as shown in Fig. 1. Actual data is not as regular as this, particularly for large values of n, but in a great number of cases the values of F_n larger than about 0.1 do come close to lying on a straight line[1], on a semilog plot, the values for higher n's are often too small to count. The plot of F_n against $G_n/\bar{n}$ is not linear. However Fig. 2 shows that for F_n greater than about 0.2, the points lie approximately on a straight semilog line, indicating that they lie roughly on the line

$$\ln F_n = \beta\bar{n}(x_n - 1) \quad \text{or} \quad F_n = \exp\left[\beta\bar{n}(x_n - 1)\right] \qquad (6)$$

where $x_n = G_n/\bar{n}$ and where $\beta\bar{n}$ is approximately 3.

The Bradford Distribution.

The Bradford distribution assumes an exponential dependence of F_n on G_n, rather than on n. If the total production $\bar{n}N_p$ is divided into equal fractions or "zones", with the "core" fraction the production of the most productive items,

the next zone that of the next most productive items and so on then, in many cases, it turns out that the number of items in the successive zones are related geometrically. To put it another way, if the unit range of $x = G_n/\bar{n}$ is divided into M equal parts, with items in the core from $x = 0$ to $x = x_1 = 1/M$, being the most productive, those in the second zone, from $x = x_1$ to $x = x_2 = 2/M$, the next most productive and so on, then the fraction of items $F(x_r) - F(x_{r-1})$ in the r'th zone bear a ratio α to those in the (r-1)'st zone, where α is independent of r. This is the same as saying that

$$F(x_r) = \alpha^{r-M} \quad \text{or} \quad \ln F(x_r) = (r-M)\ln\alpha \quad \text{or}$$
$$F(x_r) = \exp\left[\beta\bar{n}(x_r - 1)\right] = \alpha^{-M}\alpha^{r} = A\,\alpha^{r} \quad \text{where} \qquad (7)$$
$$x_r \equiv G(x_r)/\bar{n} = r/M \quad \text{and} \quad A = e^{-\beta\bar{n}} = \alpha^{-M}$$
$$F(x_M) = F(1) = 1 \quad \text{then} \quad F(x_1) = \alpha^{1-M} = \exp\left[(\tfrac{1}{M} - 1)\beta\bar{n}\right]$$

[← Factor A comes in Eq.(8); α is the Bradford ratio between zones.

Thus the Bradford distribution assumes that Eq.(6) holds exactly. Since Fig. 2 shows that the geometric distribution satisfies Eq. (6) only approximately over the upper o.8 of the range of F and not over the lower 0.2 (the range for large values of n), we must now work out the exact solution of Eqs. 6, to see to what extent and over what range it will approximate the geometrical distribution -- and, indeed, to see just how F_n must depend on n, as well as on G_n.

Leimkuhler[3] has worked out the case where the effective range of n is very large (many items with n larger than 100) so that x_r can be considered to be a continuous variable. In quite a few cases, however, the great majority

of items have productivity less than about 20. In these cases we cannot neglect the fact that the productivity n is an integer, not a continuous variable, and we must work out the distribution in productivity n required by Eqs. (7). We note one important general property, apparent from Leimkuhler's analysis; the distribution cannot extend indefinitely out to higher and higher productivities n. Since, in the limit of large n, f_n turns out[3] to be proportional to $(1/n^2)$, the cumulative production function G_n of Eqs. (4) cannot extend to infinity, because the sum of (1/n) to $n \to \infty$ diverges.

In the case of the geometric distribution the "tail" of the productivity curve of nf_n diminishes rapidly enough, for high n, so we can neglect the cumulative effect of this tail without appreciably affecting the result. But with the Bradford distribution, the tail of the productivity curve never becomes negligible, and it is necessary to say that the equation specifying the fraction f_n of items of productivity n can hold only from n = 1 out to some upper limit n = N, the outer limit of the "core", and that the distribution of the high productivity items in the core is sufficiently scattered so that all we can do is to give the fraction of items, F_N, in the core and its net production function G_N, without attempting further detail there.

Thus the Bradford distribution begins with an upper limit n = N, beyond which (in the core) we can specify only the fraction F_n and the production function G_n, but below which we can specify the fraction f_n of items with productivity n and the cumulative distribution F_n and production function G_n of Eqs. (3) and (4) (for each and every n). The effect of the core items, those with n ~~greater~~

greater than N, is given only by specifying the cumulative quantities F_N and G_N. In the lower range, however, we see, from Eqs.(6) and (7) that

$$F_n = \sum_{m=n}^{N-1} f_m + F_N = Ae^{\beta G_n} \quad \text{where} \quad G_n = \sum_{m=n}^{N-1} m f_m + G_N \quad \text{or}$$

$$Y_n = \beta F_n = \sum_{m=n}^{N-1} y_m + Y_N = \beta A e^{Z_n} \; ; \quad Z_n = \beta G_n = \sum_{m=n}^{N-1} z_m + Z_N \qquad (8)$$

$$\text{where} \quad y_n = \beta f_n \quad \text{and} \quad z_n = n y_n = \beta n f_n$$

Iterative solution of this set of equations can be started from $n = 1$ outward by noting that, since $Z_{n+1} = Z_n - z_n$,

$$y_n = Y_n - Y_{n+1} = \beta A(e^{Z_n} - e^{Z_{n+1}}) = \beta A e^{Z_n}(1 - e^{-z_n}) \quad \text{or}$$

$$n y_n = z_n = n Y_n (1 - e^{-z_n}) \; ; \quad Y_{n+1} = Y_n - (z_n/n) \qquad (9)$$

We start with an assumed value of Y_1, compute z_1 and thus y_1 by finding the value of z_1 that satisfies the first of Eqs.(9) with the assumed value of Y_1, then computing Y_2 from the second of Eqs.(9) and so on. For all but one unique value of Y_1 the computed sequence of z_n's begins to oscillate and soon negative values appear. But for the value of Y_1 approximately equal to 1.495483 the sequences of z_n and y_n diminish smoothly to zero as n increases. In accord with Leimkuhler's continuous solution, the formulas for z_n and y_n, for n large are expressible in terms of inverse powers of n. The approximate formulas

$$z_n \simeq \frac{1}{n} - \frac{0.00487}{n^2} - \frac{0.19424}{n^3} \; ; \quad y_n = z_n/n \qquad (10)$$

holds to better than one percent accuracy for n greater than 6.

Since F_1 must equal unity, β must equal Y_1 and the values of the distribution functions can be found by dividing by β. They are tabulated in Table I and plotted against n in

Fig.3. It obviously is not an exponential distribution in n. The approximate formulas for f_n and for the cumulative distributions F_n and Q_n, for n large, are

$$f_n \simeq \frac{0.668680}{n^2} - \frac{0.003256}{n^3} - \frac{0.129884}{n^4} \quad (\text{for } n > 6)$$

$$F_n \simeq \frac{0.668213}{n} + \frac{0.3578}{n^2} - \frac{0.25}{n^3} \quad (\text{for } n > 20) \qquad (11)$$

$$G_1 - G_n = Q_n \simeq 0.269302 + 0.668620 \ln n - \frac{0.335153}{n} + \frac{0.0465}{n^2} \quad (n > 15)$$

Generating a Bradford Distribution.

To generate a Bradford distribution we pick a value of n to be the maximum value N, and use the value of F_N to be the fraction of items in the core. We then arbitrarily pick a corresponding value of G_N, which must be equal to or slightly larger than NF_N -- otherwise the mean productivity of the core, G_N/F_N, will be less than the productivity N of the most productive items in the next zone, contrary to our assumptions. Values of G_n are then found by adding values of nf_n successively, or else by subtracting the quantities Q_n in Table I from $G_N + Q_N = \bar{n}$,

$$G_n = G_N + \sum_{m=n}^{N-1} m f_m = G_N + Q_N - Q_n = \bar{n} - Q_n \qquad (12)$$

where $\bar{n} - Q_N$ must be no smaller than NF_N.

We can, of course choose the distribution to correspond to a specified value of average productivity $\bar{n}$, though the choice is limited by the requirement that $\bar{n}$ should not be less than $Q_N + NF_N$. In either case

$$F_n = e^{-\beta Q_n} = A e^{\beta G_n} \quad ; \quad A = \exp\left[-\beta(G_N + Q_N)\right] = e^{-\beta \bar{n}}$$
$$\beta = 1.495483 \quad ; \quad \alpha = \exp(\beta \bar{n}/M) \qquad (13)$$

as is required by Eqs.(8).

If we desire a "standard" Bradford distribution, where the total production is divided into M equal parts, we start by choosing a value of mean productivity $\bar{n} = G_1$; then $A = e^{-\beta\bar{n}}$ to be used in Eqs.(13) for F_n. The values of F_N and G_N for the core collection are then

$$F_N = F(x_1) = \exp(-\beta\bar{n} + \beta G_N) \; ; \; G_N = G(x_1) \; ; \; x_r = r/M$$
$$\text{and} \quad F(x_r) = \exp\left[-\beta\bar{n}(1 - x_r)\right] \; ; \; G(x_r) = \bar{n}\, x_r \qquad (14)$$

As with the limit mentioned following Eqs.(12), there is an upper limit on the choice of M for a given choice of $\bar{n}$, as explained shortly.

Of course the successive values of $F(x_r)$ do not come out exactly equal to any one of the values of F_n in Table I, which means that some of the collection of items of a given productivity n will have to be divided between two zones. For example, for $\bar{n} = 3$ and $M = 4$, the fraction of items in the core collection is $F(1/4) = 0.03457$, which is just less than the value of F_n in Table I for $n = 19$. The value of $G(1/4)$ for the core collection is , from Eqs.(14), 3/4, so the mean productivity for the core is $\left[G(1/4)/F(1/4)\right] = 21.7$, which is greater than the productivity $n = 19$ of the first items in the second zone, $r = 2$. On the other hand, if we try to divide the $\bar{n} = 3$ distribution into $M = 5$ parts, $G(1/5) = 0.6$, $F(1/5) = 0.02762$, which is just less than the value of F_n for $n = 25$. But the net productivity of the core, $\left[G(1/5)/F(1/5)\right] = 21.8$. This is less than the productivity of the first items in the second zone, 25, which violates our assumption that the core items have the largest productivity.

Bradford[2] discussed a case for $M = 3$, for which he estimated α to be about 5. For this case, according to Eq. (13), $\bar{n} = (M/\beta)\ln\alpha \simeq 3.23$. Taking $M = 3$ and $\bar{n} = 3.23$, Eq.(14) shows that $F_N \equiv F(1/3) = 0.03994$, practically equal to F_{17} of Table I. Thus $N = 17$, i.e., the core contains all items of productivity greater than 17. The mean productivity of the core, $[G(1/3)/F(1/3)] = [\bar{n}/3F(1/3)] = 26.9$. In other words the productivity of the core is considerably greater than $N = 17$; evidently there are individual items in the core with productivities from 18 up to well above 27.

Actually, for $M = 3$, $\bar{n}$ can be as small as 2.2433, in which case $F(1/3) = 0.10683$. Since the next largest value of F_n in Table I is $F_6 = 0.12092$, N is 6, with the mean productivity of the core being $[\bar{n}/3F(1/3)] = 6.9997 \simeq 7$. In this case the core consists entirely of items of productivity 7; the other two zones divide the items of productivity 1 to 6 between them in a manner shortly to be taken up.

An illustration of these matters is shown in Fig.4, for the case of $\bar{n} = 2.5$ and $M = 4$. The fraction of items in the core collection, F(1/4), is 0.061, a little larger than F_n for $n = 11$. The value of G(1/4) for this case is 2.5/4 = 0.625, so the net productivity of the core, G(1/4)/F(1/4), is also about 11. Thus this example is nearly the smallest value of $\bar{n}$ for which a division into 4 zones is possible. The abscissa of the figure is $x = G(x)/\bar{n}$. The large steps show the values of F(x) for each of the 4 zones, ending with F = 1 for the last step. The small steps give the values of F_n, from Table I, against G_n. These values also fall on the same dashed, straight line, on the semi-log plot, as do the large steps, demonstrating the geometric dependence of F on G, as required by the Bradford distribution equations (6) and (7).

Note that, if one were fitting actual data, for which the mean productivity $\bar{n}$ was 2.5, into 4 zones, the items for $n = 11$ would have to be divided between the core and the second zone, the items for $n = 5$ would have to be divided between the second and the third zone and the items for $n = 2$ would have to be divided between the third and the fourth zone. Note also that the division of the items in each of these subzones will have to be allocated harmonically, not linearly. For example, if $G(x_r)$ comes midway between G_{n+1} and G_n, then the fraction $\sqrt{F_n F_{n+1}} - F_{n+1}$ of $f_n = F_n - F_{n+1}$ is allocated to the lower zone and the fraction $F_n - \sqrt{F_n F_{n+1}}$ is allocated to the upper zone. In the more general case, when

$\frac{G(x_r) - G_{n+1}}{G_n - G_{n+1}} = \mu$, where $G(x_r) = \bar{n}(r/M)$ lies between G_n and G_{n+1} and where all the $N_p(F_n - F_{n+1})$ items have productivity n, then $N_p(F_n^{\mu}F_{n+1}^{1-\mu} - F_{n+1})$ (15)

of these items are allocated to the r'th zone and $N_p(F_n - F_n^{\mu}F_{n+1}^{1-\mu})$ are allocated to the (r+1)'st zone

Comparison of the Distributions

Actual data, of course, fits neither of these two distributions exactly and, since there is not much difference between the two in mid-range, it may be difficult to distinguish which fits the data best. Plotting the values of the cumulative distribution F_n against n will help; if the plot, for the first 5 or 6 values of n, is nearly a straight line on semi-log paper, the geometric distribution is the better fit; if the plot is concave upward, like that of Fig.3, the Bradford distribution is better (if the plot is markedly concave downward then neither distribution fits well). One can, of course, plot F against G on a semilog plot, as in Figs.2 and 4, though this plot does not distinguish as clearly between the distributions. If the vertices of the steps tend to form a straight line, the Bradford distribution is better; if they are concave downward, as in Fig.2, the geometric distribution may be a better fit.

However, in many cases the planning factors to be obtained by the use of one or the other distribution need not be very precise in order to guide a policy decision. If so then the choice of the distribution to use depends on the nature of the conclusions to be drawn. The advantage of the Bradford

distribution -- and the Bradford way of organizing the data -- is perhaps the ease of counting the items. However the distribution is less flexible; the value of β is fixed; both $\bar{n}$ and M must be specified and there is an upper limit to the choice of M. The conclusions that can be drawn from the distribution are somewhat limited and if one wishes to go beyond concluding that the fraction (r/M) of the total production of the collection comes from the fraction F(r/M) of the items, the further conclusions must be worked out numerically, using the numbers in Table I.

On the other hand, the geometric distribution can be manipulated analytically, once the value of the single parameter $\gamma = 1 - (1/\bar{n})$ is computed for the collection. Because the distribution of Eqs.(5) can be expressed in symbols, rather than having to be given in numerical form, the related properties, such as variances, characteristics of different parts of the collection and so on, can be expressed as explicit functions of γ, rather than having to be evaluated numerically for each value of $\bar{n}$ and M, as is necessary for the Bradford distribution. Furthermore, if one wishes to combine it with other distributions, as one does when calculating predicted changes with time[1], at least part of the calculation can be made symbolically with the geometric distribution, rather than having to work it out numerically for each value of the parameters.

Nevertheless both distributions have their value in describing the properties of a collection of items such as books or periodicals or other informational units.

TABLE I

The Bradford Distribution.

n	F_n	f_n	nf_n	Q_n
1	1.00000	0.57986	0.57986	0.00000
2	.42014	.15903	.31806	.57986
3	.26110	.07250	.21750	.89793
4	.18861	.04118	.16473	1.11542
5	.14742	.02650	.13250	1.28015
6	.12092	.01844	.11067	1.41265
7	.10248	.01358	.09508	1.52332
8	.08890	.01041	.08328	1.61840
9	.07849	.00823	.07408	1.70168
10	.07026	.00667	.06671	1.77576
11	.06358	.00516	.06067	1.84247
12	.05807	.00464	.05563	1.90313
13	.05343	.00395	.05136	1.95876
14	.04948	.00341	.04770	2.01012
15	.04608	.00297	.04453	2.05782
16	.04311	.00261	.04175	2.10234
17	.04050	.00231	.03930	2.14409
18	.03819	.00206	.03712	2.18339
19	.03613	.00185	.03517	2.22050
20	.03427	.00167	.03341	2.25567

$$F_n = \sum_{m=n}^{\infty} f_m \qquad Q_n = \sum_{m=1}^{n-1} m f_m$$

References.

[1]Morse, P.M., Library Effectiveness, Chapters 2 and 5, Cambridge, Mass., The MIT Press, 1968.

[2]Bradford, S.C., Documentation, London, Crosby Lockwood, 1948.

[3]Leimkuhler, F.F., The Bradford Distribution, Journal of Documentation, Vol.23, No.3, p.197-207, 1967.

Figure Captions.

Fig.1. Examples of the geometric distribution. Cumulative fractions of items F_n versus productivity n,

Fig.2. The geometric distribution F_n versus cumulative production factor G_n.

Fig.3. The Bradford distribution. Cumulative fraction of items F_n versus productivity n.

Fig.4. The Bradford distribution for mean productivity $\bar{n} = 2.5$. Small steps show F_n divided into integral productivity steps; Large steps show $F(x_r)$ in equal steps of production factor G.

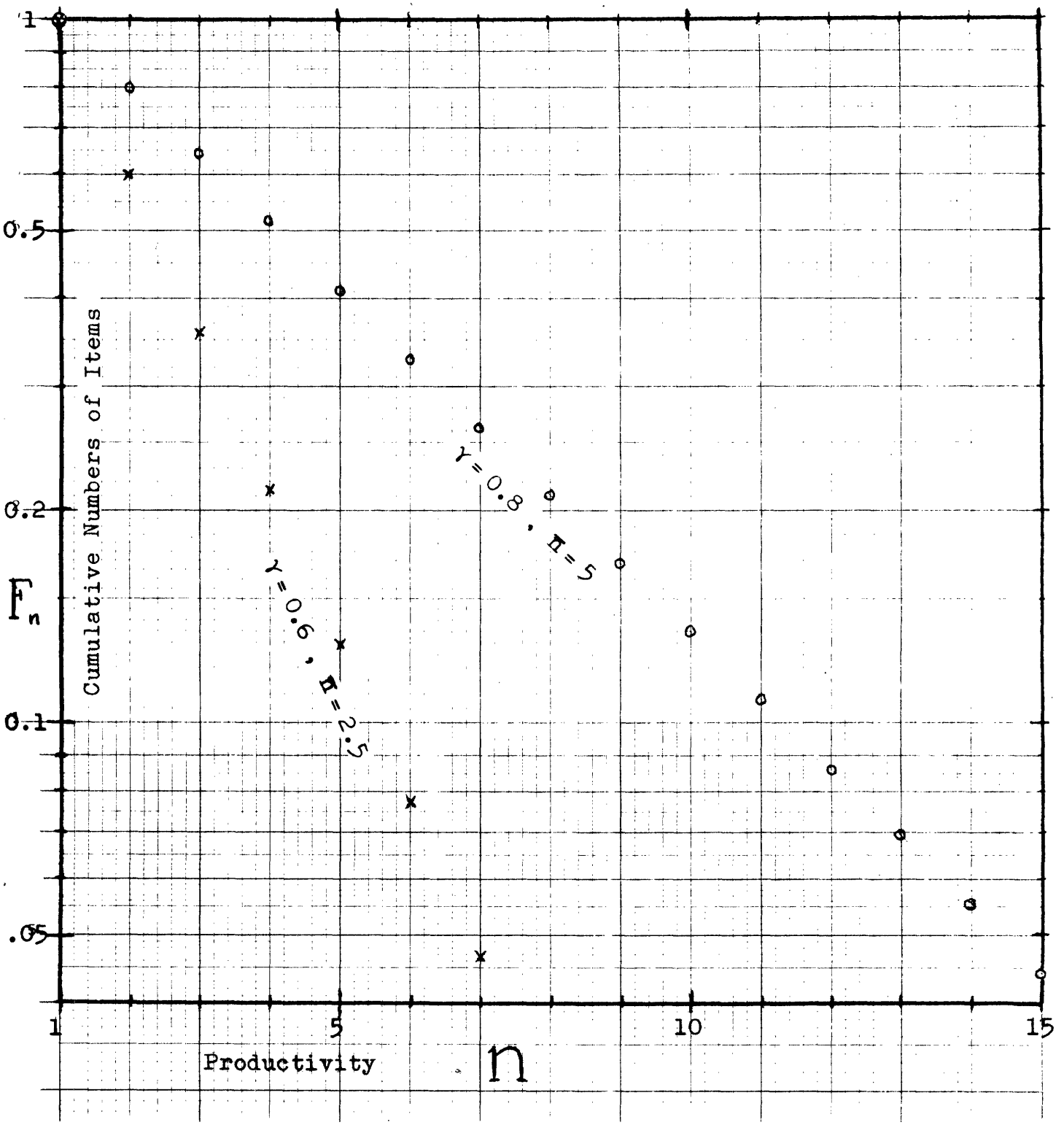


Fig.1. Examples of the geometric distribution. Cumulative fractions of items F_n versus productivity n.

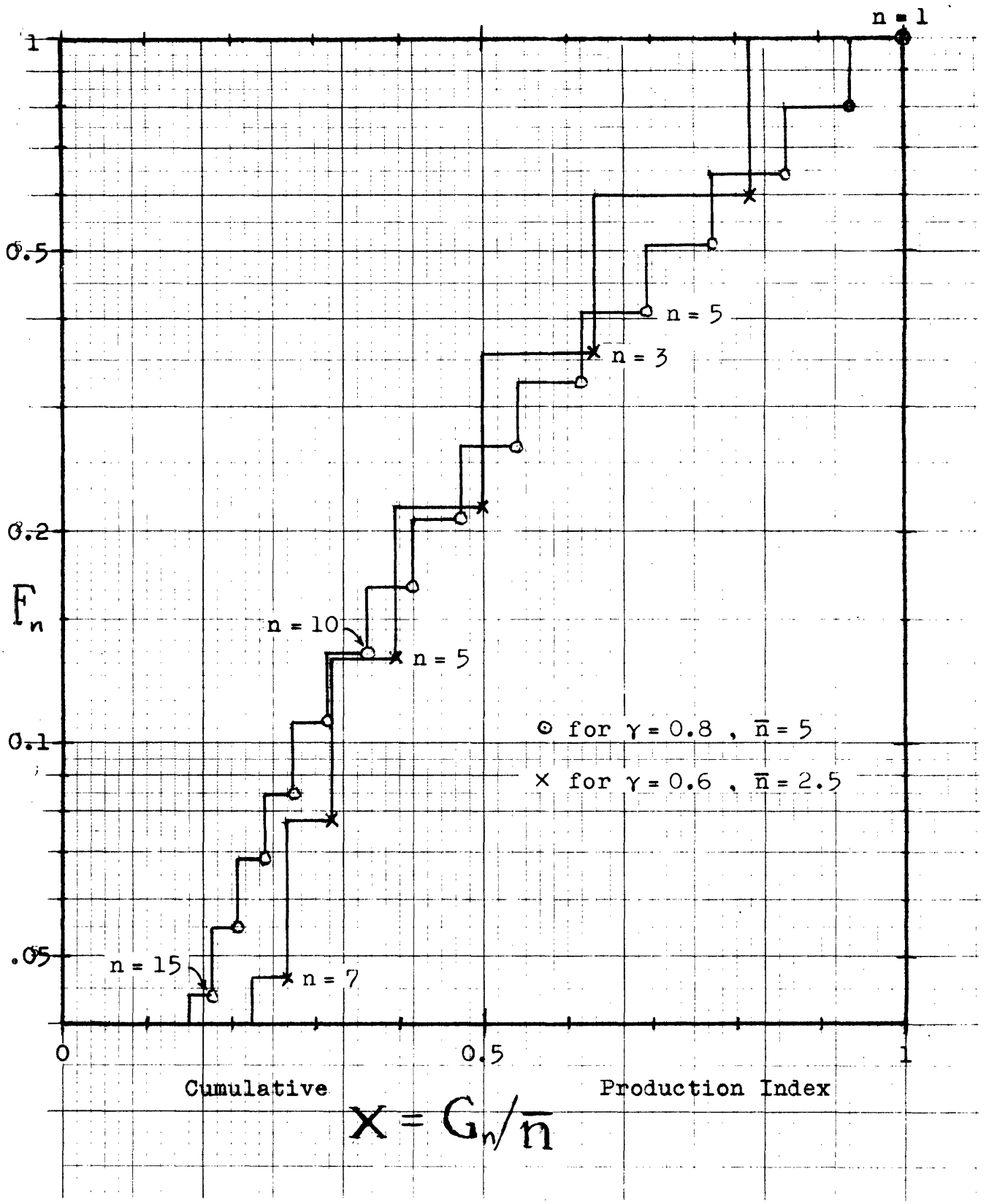


Fig.2. The geometric distribution F_n versus cumulative production factor G_n.

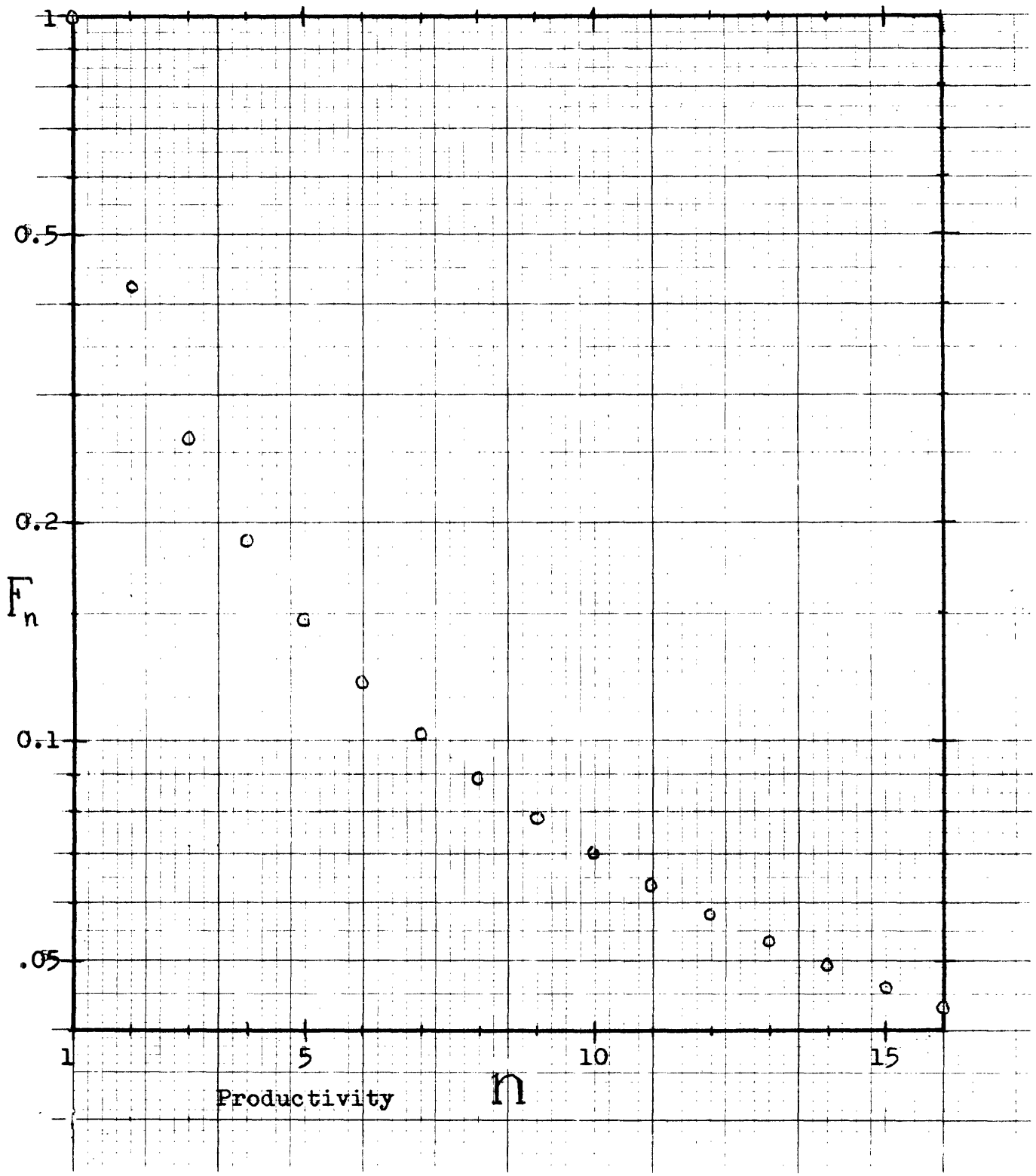


Fig.3. The Bradford distribution. Cumulative fraction of items F_n versus productivity n.

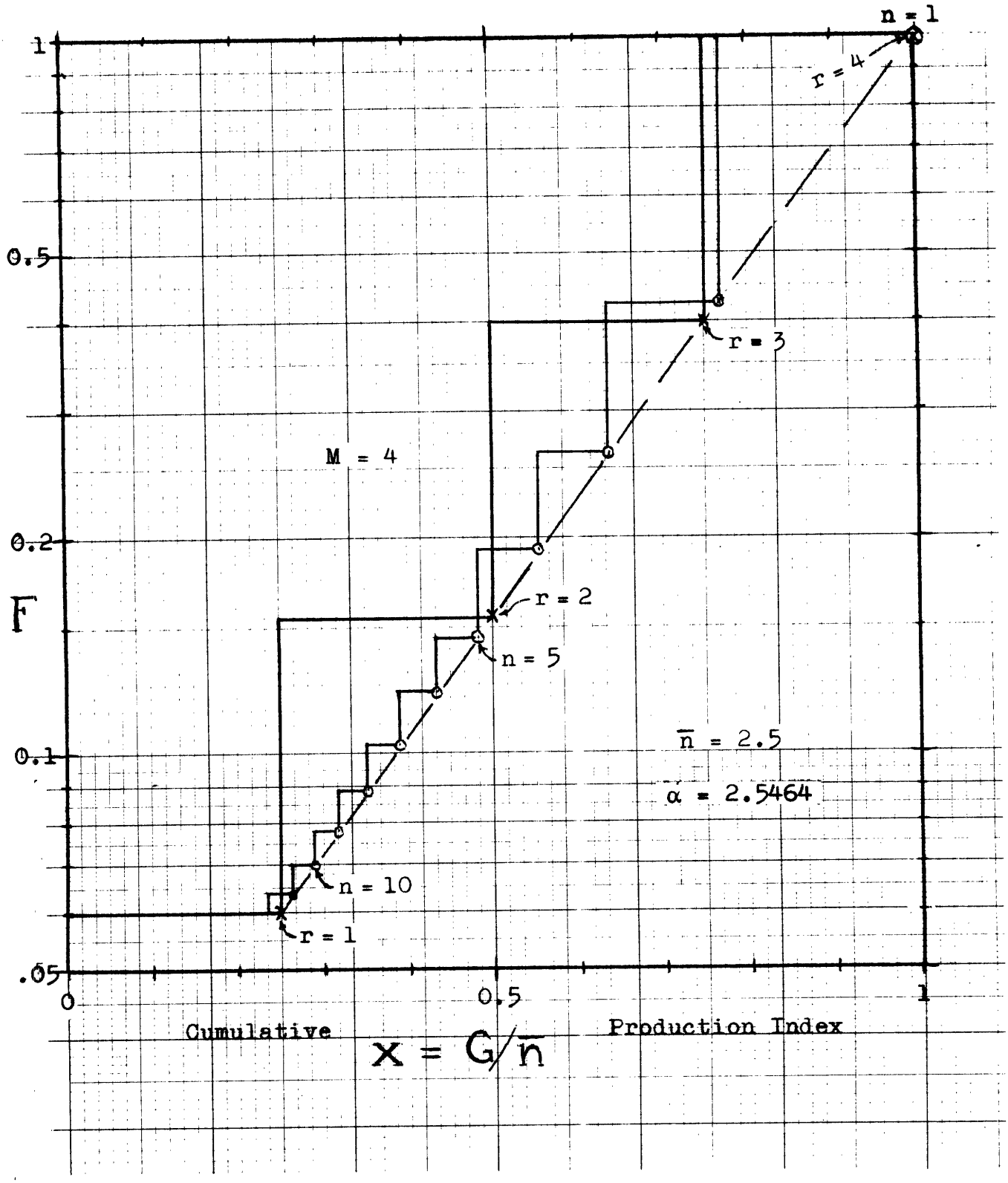


Fig.4. The Bradford distribution for mean productivity $\bar{n} = 2.5$. Small steps show F_n divided into integral productivity steps; Large steps show $F(x_r)$ in equal steps of production factor G.