SOME FUNDAMENTAL CURVES FOR THE SOLUTION OF SAMPLING PROBLEMS

By Edward C. Molina

East Orange, N. J.

1. Summary. In using collateral information in an inverse probability situation to estimate a population fraction from a sample fraction it is necessary to use some particular form for the *a priori* probability function. This paper points out the advantages of using $Kx^r(1 - x)^s$ for this purpose. The application then involves only the Incomplete Beta Function.

Graphs of the 10, 25, 50, 75 and 90 per cent points of the Incomplete Beta Function are given. They cover a range which includes and extends previous tabulations.

2. Introduction. The engineer, scientist or industrialist is often confronted with the following "sampling" problem:

> "The probability, p, of an event happening in a single trial is constant from trial to trial, but the numerical value of this constant is unknown. A series of n trials is made and the event happens c times, $c \leq n$. What light does this statistical data shed on the unknown value of p?"

As a concrete example, suppose that a new type of brakes is proposed for a given class of steam locomotives making the run from Buffalo to Detroit.[1] Let each of 30 locomotives be equipped with a set of the new brakes and given a trial run. Of these, 26 make satisfactory runs, so far as the behavior of the brakes is concerned; the remaining four encounter difficulties. Here, the event of interest is a satisfactory run, $n = 30$ and $c = 26$. What "weight" (confidence[2]) may the design engineer assign to the assumption that, say, $25/30 \leq p \leq 27/30$?

Practical decisions involving such statistical data are usually based on a combination of the data with "collateral" information. In fact, the applied statistician is all too familiar with the extreme case where the statistical data are so meagre as to provide no information and where a decision must be made *now*—in these cases the decision is made solely on the basis of the collateral information, and rightly so.

The methods of statistical analysis and presentation developed up to the present have concentrated on the other extreme case, where the statistical data are so good that collateral information can be neglected.

[1] This fictitious example convicts the writer of total ignorance of railroad engineering. Nevertheless, the illustration brings out, in concrete terms, the class of sampling problems under consideration.

[2] The purely intuitive meaning to be attached to "weight" and "confidence" is the same. However, the curves presented with this paper are not based on the theory which underlies what are known, in statistical literature, as "confidence intervals".

There is a real need for methods of analysis and presentation to be used where both the statistical data and the collateral information should be used. However, when the significance of the collateral information is adequately expressed by a function $w(x)$, x being a permissible value of the unknown p, the classic Bayes-Laplace theory (see [1]) of inverse probability gives the solution to a sampling problem.

The purpose of this paper is to present a set of sampling curves based on a $w(x)$ function whose form embodies some important properties.[3]

3. Hardy's collateral frequency function. Consider again the locomotive brakes problem. The new design may have been carefully engineered, in accordance with well-known principles, to reduce costs at the expense of a slight reduction in reliability of operation. In such a situation, the collateral information would be somewhat as follows: There is a high "probability" that the unknown value of p is a little below the known value for the old type of brakes. Moreover, it may be assumed that the "probability" drops rapidly for values of p departing materially from this old value. Suppose the latter is $p = .95$; then the collateral information would be presented by some such curve as number 5 in Figure 1, the mode (peak) of this curve being at .90, which is slightly below the old .95 value.

Number 5, of Figure 1, belongs to the family of curves corresponding to the frequency function

$$w(x) = Kx^r(1 - x)^s$$

This form for $w(x)$ was suggested, in 1889, by the British actuary Sir George F. Hardy (see [2]) for the construction of mortality tables. Its mode, mean and variance are given by the equations

$$\begin{aligned}
\text{Mode} &= r/(r + s)\\
\text{Mean} &= (r + 1)/(r + s + 2)\\
\text{Variance} &= (r + 1)(s + 1)/(r + s + 2)^2(r + s + 3)
\end{aligned}$$

G. J. Lidstone (see [3]) has pointed out that the Hardy form for $w(x)$ has two important advantages: First—"By suitable choice of r and s any required values of the mode or mean and the variance of z_x can be reproduced, and thus a great variety of distributions may be approximately represented." Lidstone's z_x is our $w(x)$. Second—"The factors x^r and $(1 - x)^s$ unite in the simplest and most elegant way with similar factors in the Laplacian integrand . . . ".

[3] Many statisticians, including a referee of this paper, feel that it is a common situation to have the collateral information so vague and elusive that it is virtually impossible to take it into account via inverse probability. (The author doubts this.) Such statisticians may wish to use the Clopper-Pearson confidence intervals, using no collateral information, in which case these curves can be used as indicated by Scheffé ("Note on the use of the tables of percentage points of the incomplete beta function to calculate small sample confidence intervals for a binomial p", *Biometrika*, August, 1944).

From this second advantage there follows a third which will be presented in section 6 below.

4. Theory. The Bayes-Laplacian formula gives us

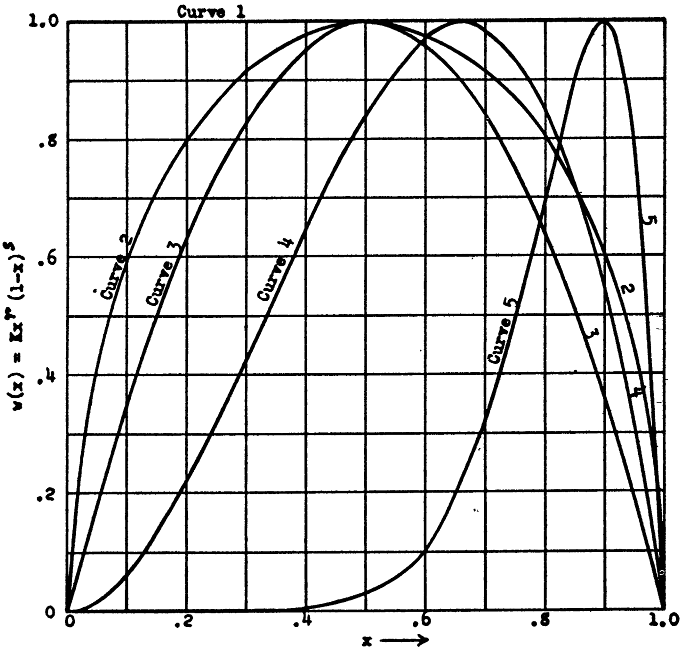
$$(1) \qquad P(p \leqq X) = \int_0^X w(x)x^c(1-x)^{n-c}\,dx \Big/ \int_0^1 w(x)x^c(1-x)^{n-c}\,dx$$

for the "a posteriori probability" that $p \leq X$. In this formula, the product

FIG. 1

Particular forms of the a priori (collateral information) function:

$$w\,(x) = K\,x^r\,(1-x)^s$$



Curve	r	s	Form
1	0	0	K
2	$\frac{1}{2}$	$\frac{1}{2}$	$Kx^{\frac{1}{2}}(1-x)^{\frac{1}{2}}$
3	1	1	$Kx(1-x)$
4	2	1	$Kx^2(1-x)$
5	9	1	$Kx^9(1-x)$

$x^c(1-x)^{n-c}$ takes care of the fact that the event happened c times in the n trials; the factor $w(x)$ represents, quantitatively, the collateral information.

Adopting, now, Hardy's frequency function, we assume that

$$(2) \qquad w(x) = Kx^r(1-x)^s,$$

r and s being assigned values in accordance with the collateral information pertaining to the particular problem under consideration. Theoretically, the constant K should be such that

$$\int_0^1 w(x)\,dx = 1,$$

but, since $w(x)$ enters in both numerator and denominator of (1), any desirable value may be given to K. Advantage has been taken of this in constructing Figure 1; to facilitate comparison of the five curves shown therein, for each curve K is such that the maximum ordinate is equal to 1.

The second advantage, pointed out by Lidstone, of the form adopted in this paper for the function $w(x)$ becomes apparent immediately on substitution of (2) in (1). We obtain

$$(3) \qquad P(p \le X) = \int_0^X x^C(1-x)^{N-C}\,dx \Big/ \int_0^1 x^C(1-x)^{N-C}\,dx$$

with $C = c + r$ and $N = n + r + s$. Therefore, a *single family of fundamental curves*, plotted with reference to C and N, will give the solutions for a multitude of different practical problems. To solve a particular problem, for which the values of n, c, r and s are specified, we merely enter the curves with $C = c + r$ and $N = n + r + s$. These linear relations transform all a posteriori curves, published on the assumption that $w(x)$ is a constant, into fundamental curves; namely, that they are applicable with the more general form (2). For example: The information given on the sheets of inverse curves (inserted in the back cover pocket) of Col. Leslie E. Simon's *Engineer's Manual of Statistical Methods* includes the restriction "that prior to sampling, one lot fraction defective is as likely as another". It is now obvious that the use of Col. Simon's curves is not so limited; his curves may be used in any situation wherein the available collateral information is covered by the assumption that $w(x)$ has the Hardy form. Likewise, the "Weight = .98" and "Weight = .8" curves ("confidence", in the intuitive sense), presented by R. P. Crowell and the writer in their paper now have a much wider range of applicability.

5. Curves. The ratio of definite integrals in equation (3) is tabulated, in a different notation, in "Tables of the Incomplete Beta Functions", edited by Karl Pearson.

This paper	Pearson Tables	Thompson Tables (see [5])
C	$p - 1$	$(v_2 - 2)/2$
$N - C$	$q - 1$	$(v_1 - 2)/2$
X	x	tabulated value
$P(p \le X)$	tabulated value	caption to Table

The range of values of C and $(N - C)$ covered by the Pearson Tables is indicated by the shaded area in Figure 7. For curve points falling outside this range (except for $C = 1$ and 2, found from the binomial summation by trial and error) recourse was had to a series developed by the writer for the solution of some problems confronting him, as Switching Theory Engineer, in the Bell Telephone Laboratories. Many points of the $C = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12$ and 14 curves can be obtained directly from the Thompson Tables. They do not, however, give any points for the $C = 16, 18, 20, 25, 30, 40, 45$ and 50 curves. It may be added that, except for certain marginal values, the Thompson Tables were also derived from the Pearson Tables.

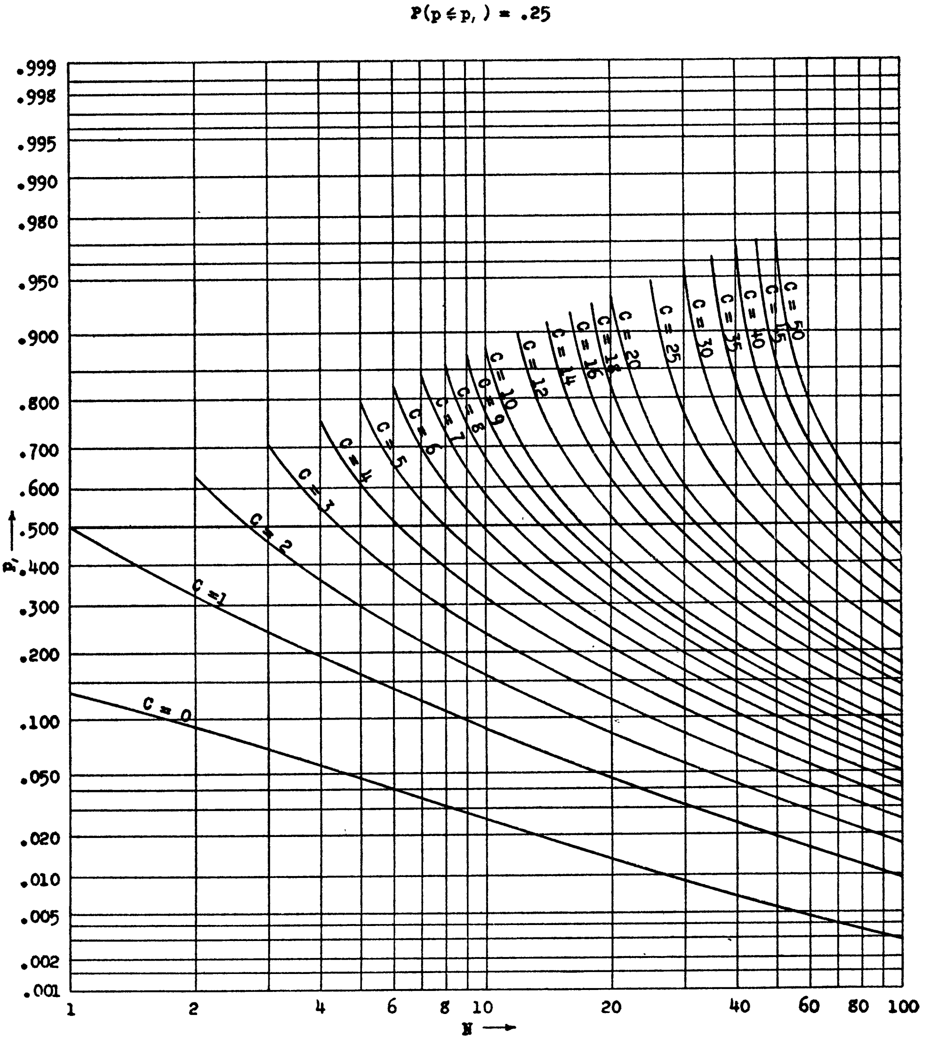


FIG. 2

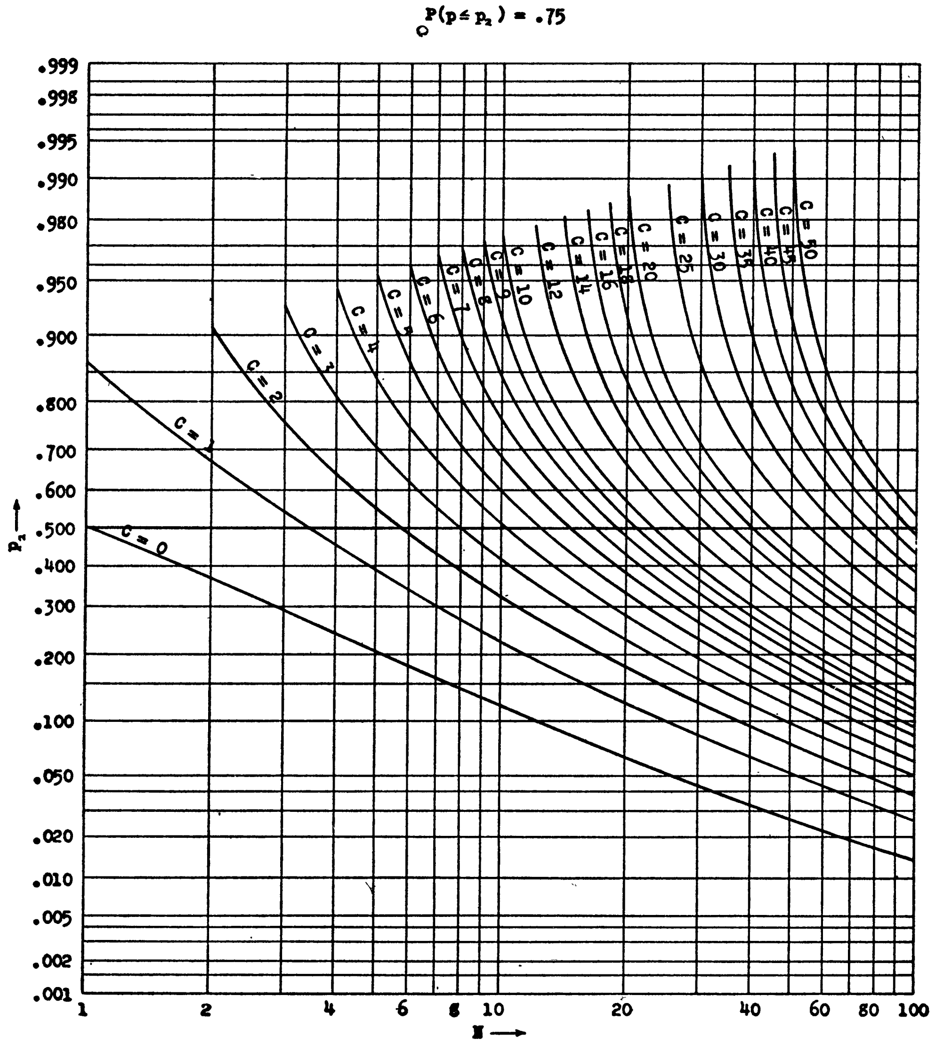


FIG. 3

Five sets of fundamental curves are submitted, namely,

Figure 2, $P(p \leq X) = .25, \quad X = p_1$
" 3, " $= .75, \quad X = p_2$
" 4, " $= .10, \quad X = p_1$
" 5, " $= .90, \quad X = p_2$
" 6, " $= .50, \quad X = p_0$

It will be noted that p_1 has been written instead of X for the curves such that

$P\ (p \leq X)$ is *less* than .50; likewise, p_2 for X for those corresponding to $P\ (p \leq X)$ *greater* than .50; p_0 for X for the $P\ (p \leq X) = .50$ curves.

P(p≤p₁) = .10

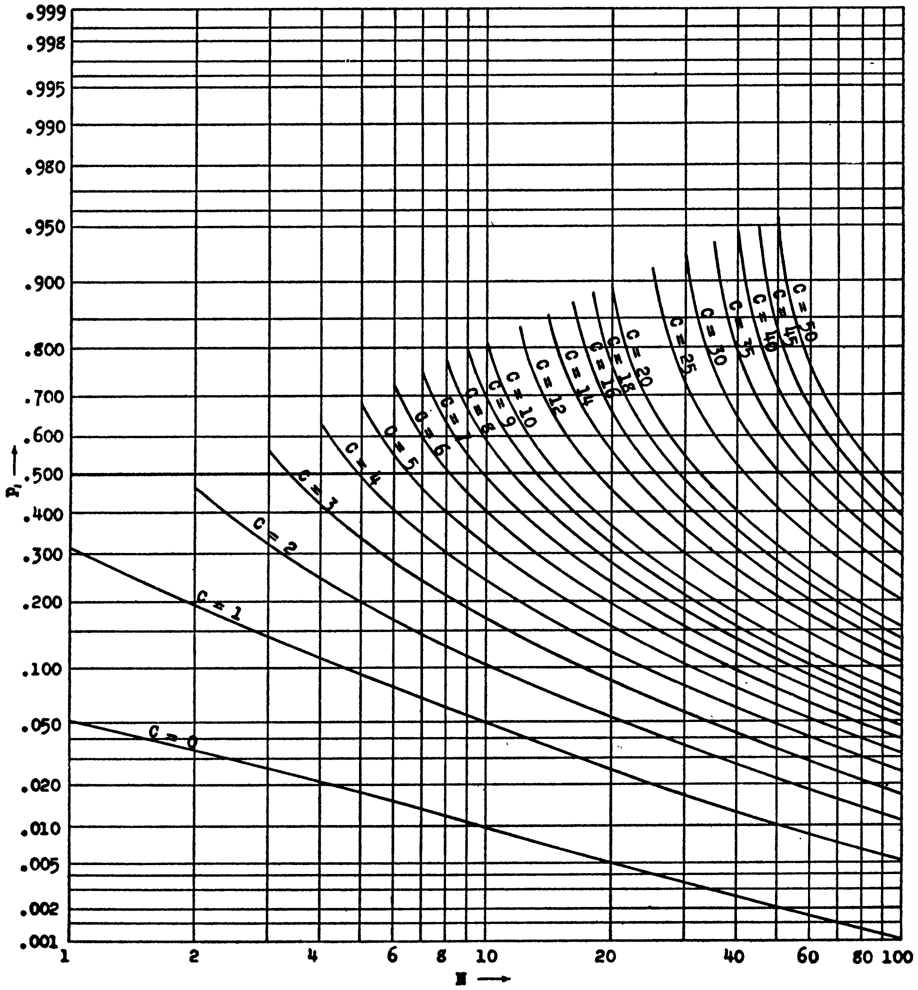


FIG. 4

For each pair of values of C and N, the curves of Figures 2 and 3 give the range

$$P(p_1 \leq p \leq p_2) = .50$$

whereas, the curves of Figures 4 and 5 give the range

$$P(p_1 \leq p \leq p_2) = .80$$

As an example of the applicability of the fundamental curves, let us reconsider the locomotive problem for which $n = 30$ and $c = 26$. It was suggested that

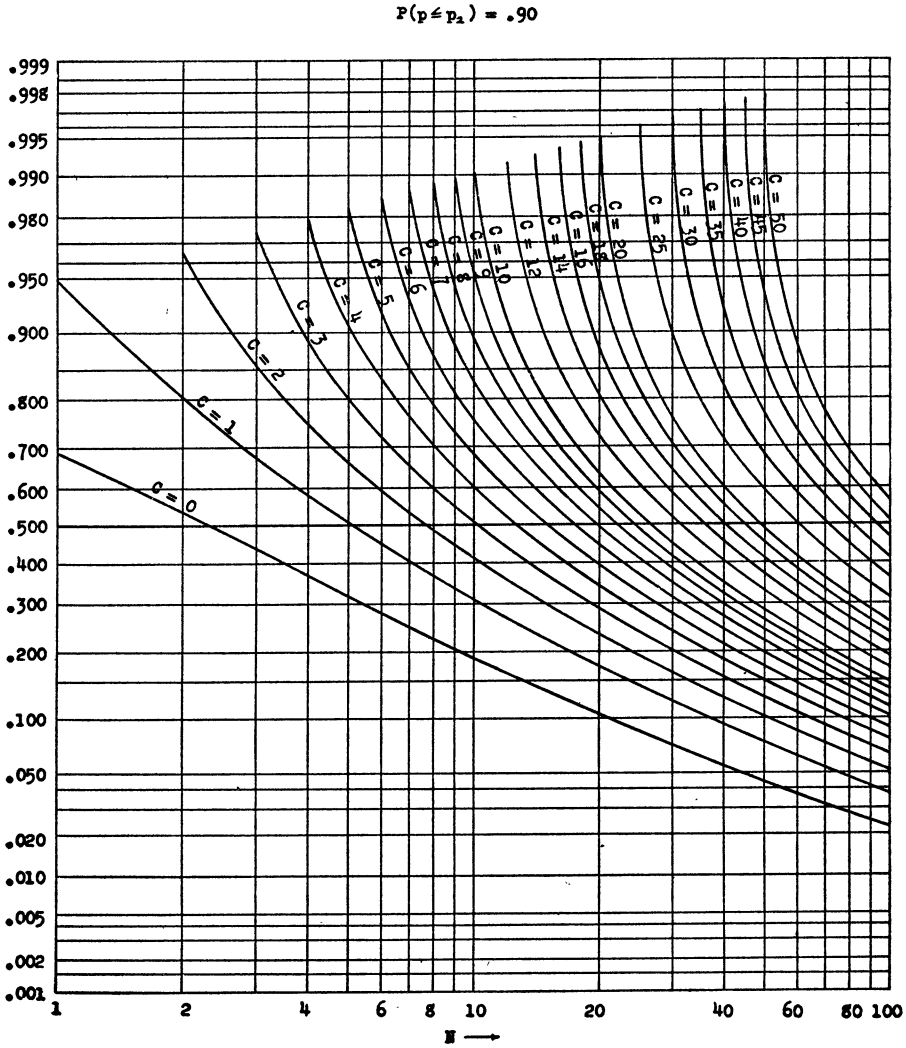


Fig. 5

the $r = 9$, $s = 1$ curve of Figure 1 might well represent the collateral information available. Therefore we take $N = 30 + 9 + 1 = 40$ and $C = 26 + 9 = 35$. Entering Figures 2, 3, 4 and 5 with this data we find

Fig.	$P(p \leq p_1)$	p_1		Fig.	$P(p \leq p_2)$	p_2
2	.25	.83		3	.75	.89
4	.10	.79		5	.90	.92

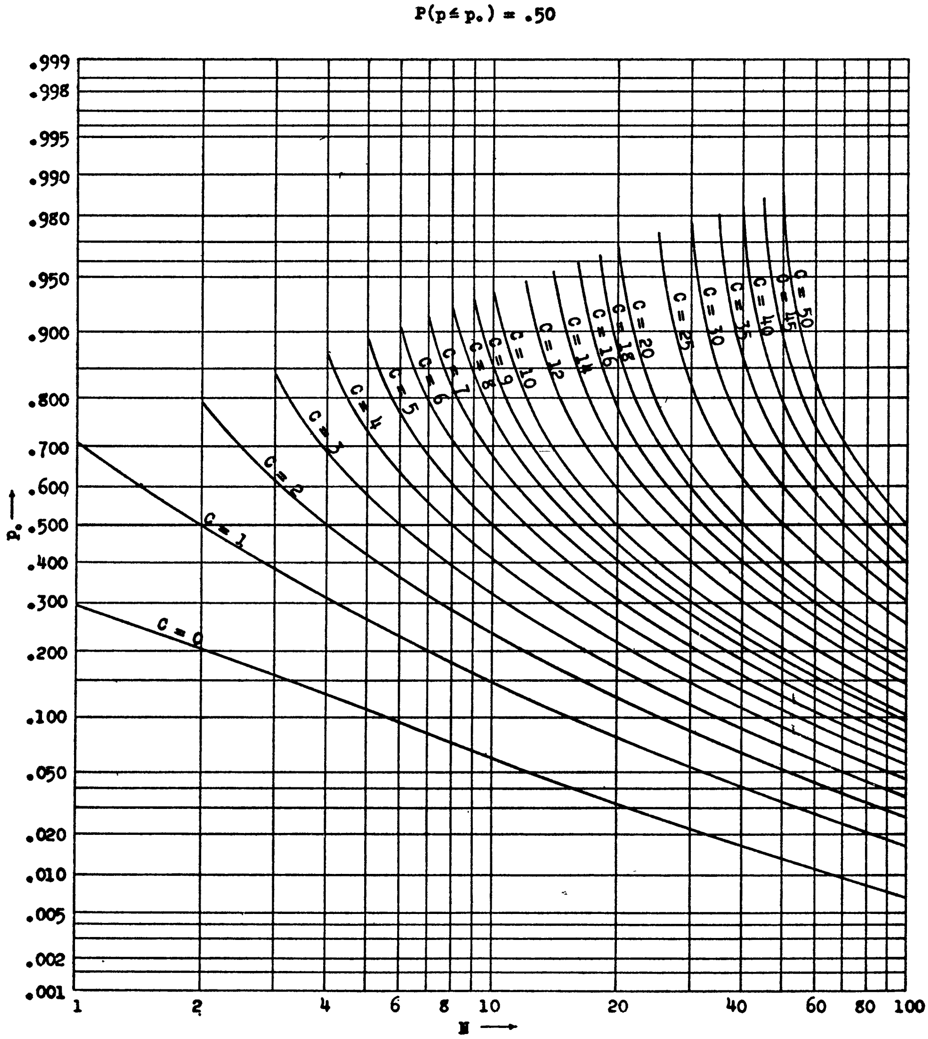


Fig. 6

Thus we have, for the unknown probability of a successful run with a new set of brakes,

$$.83 < p \leq .89, \quad \text{with weight } .50$$

and

$$.79 < p \leq .92, \quad \text{with weight } .80$$

6. Sequential property of the curves. The original draft of this paper was submitted to Dr. W. V. Houston[4] in connection with the solution of a problem

[4] Of the California Institute of Technology and now President of Rice Institute, Houston, Texas. It was Dr. Houston who gave the impetus to the publication of this paper.

"Tables of The Incomplete Beta-Function," edited by Karl Pearson, can be used for evaluation of

$$P = \frac{\int_0^P x^C(1-x)^{N-C}\,dx}{\int_0^1 x^C(1-x)^{N-C}\,dx}$$

only when values of $(N - C)$ and C are in ▨.

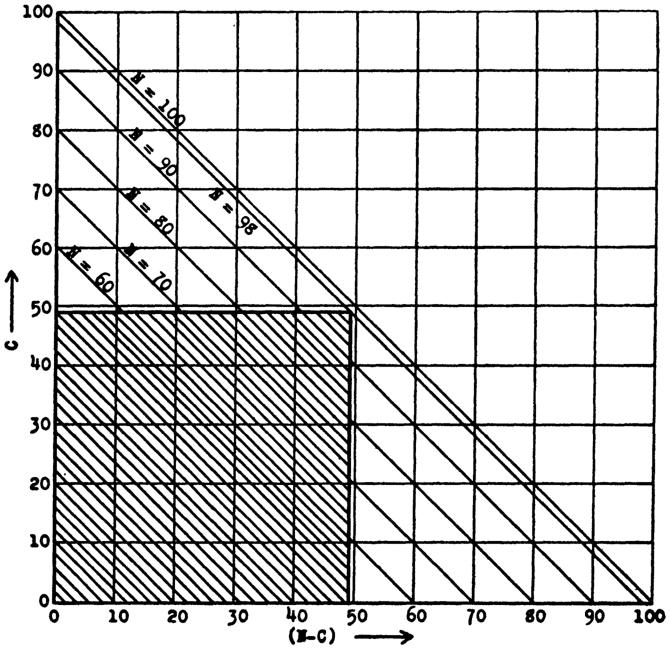


FIG. 7

in which he was interested. Regarding equation (3), Dr. Houston made a very significant comment, the burden of which may be stated as follows: Suppose that before the series of n trials had been made, it was known that, at some *earlier time*, a series of $r + s$ trials had resulted in r successful outcomes. Suppose, moreover, that the collateral information called for the assumption that, a priori, all values of p were equally likely. Under these circumstances equation (3), derived by substitution of (2) in (1), gives $P(p \leq X)$ for *two consecutive series of trials*, one of $r + s$ with r successes followed by another of n with c successes. An immediate generalization of Dr. Houston's thought shows that the fundamental curves may be entered with

$$N = n_1 + n_2 + \cdots + n_i + \cdots + n_m + r + s,$$

$$C = c_1 + c_2 + \cdots + c_i + \cdots + c_m + r,$$

for the solution of a problem involving m *consecutive series of trials*, n_i and c_i being the number of trials and successes, respectively, in the ith series; the introduction of r and s removing the restriction that all values of p were a priori equally likely.

REFERENCES

[1] T. C. Fry, "A mathematical theory of rational inference", *Scripta Mathematica*, Vol. 2 (1934).

[2] G. F. Hardy, *Trans. Faculty of Actuaries*, Vol. 8 (1920), p. 181.

[3] G. J. Lidstone, "Laplace's antecedent—probability function", *Math. Gazette*, Vol. 25 (1941), p. 162.

[4] R. P. Crowell and E. C. Molina, "Deviation of random samples from average conditions and significance to traffic men", *Bell System Tech. Jour.*, Vol. 3 (1924).

[5] Catherine M. Thompson, "Tables of percentage points of the incomplete beta-function", *Biometrika*, Vol. 32 (1941).