THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL.

By C. J. CLOPPER, B.Sc., and E. S. PEARSON, D.Sc.

(1) *General Discussion.*

In facing the problem of statistical estimation it may often be desirable to obtain from a random sample a single estimate, say a, of the value of an unknown parameter, α, in the population sampled. It has always, however, been realised that this single value is of little use unless associated with a measure of its reliability and the traditional practice has been to give with a its probable error (or more recently its standard error), in the form

$$a \pm p.e(a) \qquad \text{.................................(1).}$$

From this information it was possible, if the sample was not too small, to draw the conclusion that the unknown value of α lay within the limits

$$\alpha_1 = a - 3 \times p.e(a) \quad \text{and} \quad \alpha_2 = a + 3 \times p.e(a) \qquad \text{..............(2)}$$

with a high degree of probability. But it was neither easy to give any precise definition of this measure of probability nor to assess the extent of error involved in estimating the value of $p.e(a)$ from the sample.

The recent work of R. A. Fisher introducing the conception of the fiducial interval has made it possible under certain conditions to treat this problem of estimation in a simple yet powerful manner*. It is proposed in the present paper to illustrate on the following problem the ideas involved in this method of approach.

A sample of n units is randomly drawn from a very large population in which the proportion of units bearing a certain character, A, is p. In the sample x individuals bear the character A and $n-x$ do not. p is unknown and the problem is to obtain limits p_1 and p_2 such that we may feel with a given degree of confidence that

$$p_1 < p < p_2 \qquad \text{.................................(3).}$$

In the first place, how is this degree of confidence to be defined? The underlying conception involved in all problems of this type is extremely simple. In our statistical experience it is likely that we shall meet many values of n and of x; a rule must be laid down for determining p_1 and p_2 given n and x. Our confidence that p lies within the interval (p_1, p_2) will depend upon the proportion of times that this prediction is correct in the long run of statistical experience, and this

* R. A. Fisher, *Proc. Camb. Phil. Soc.* 26 (1930), p. 528; *Proc. Roy. Soc.* A 139 (1933), p. 343. References to the discussion of these concepts in lectures may also be found in papers published by students of J. Neyman. See for instance pp. 28—29 of a paper by W. Pytkowski written in 1929—30 and published at Warsaw in 1932, entitled, "The dependence of the income in small farms upon their area, the outlay and the capital invested in cows."

may be termed the *confidence coefficient.* Thus subject to certain approximations discussed below, arising from the fact that x can assume only discrete integral values in this particular problem, it is possible to choose the fiducial or confidence limits p_1 and p_2 in such a manner that, for example, the prediction

(1) will be correct in 95 °/₀ of cases met with in the long run of experience, and wrong in 5 °/₀, in 2·5 °/₀ because $p \leqslant p_1$, and 2·5 °/₀ because $p \geqslant p_2$.

Or again,

(2) will be correct in 99 °/₀ of cases and wrong in 1 °/₀, in 0·5 °/₀ because $p \leqslant p_1$, and in 0·5 °/₀ because $p \geqslant p_2$.

These intervals (p_1, p_2) may be termed either the central* confidence or

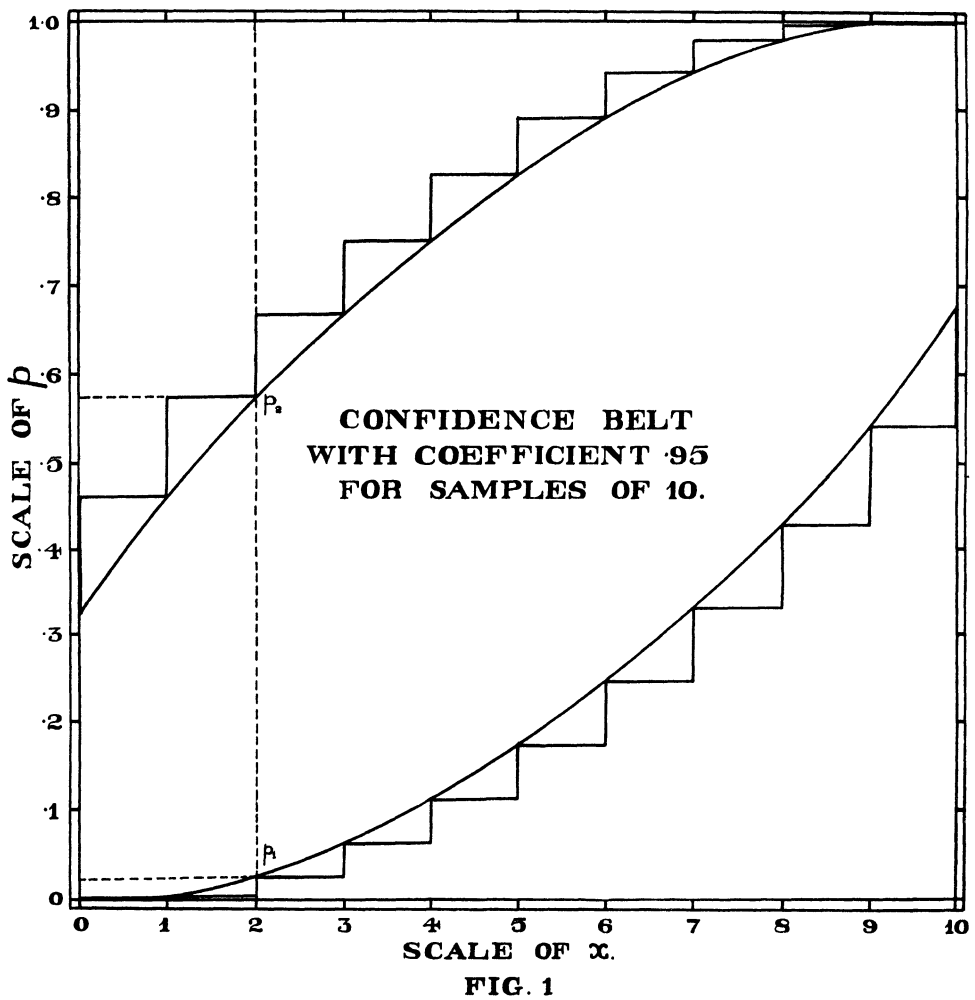

FIG. 1

* In the charts described below the coefficients ·95 and ·99 were chosen as giving two useful pairs of limits. It is not essential that the intervals chosen should be "central," but for many purposes this appears to be the most convenient arrangement.

central fiducial intervals and are associated with confidence coefficients of ·95 and ·99 respectively. In his development of the subject, R. A. Fisher has used the term "fiducial probability" to describe the chance that in the long run a correct prediction will be made of the limits within which the unknown parameter falls. The concept of fiducial probability cannot, it appears, be distinguished from that of ordinary probability, and it seems possible that the use of this term may lead to some misunderstanding, especially when associated with a "fiducial distribution." We are inclined therefore to adopt the terminology suggested by J. Neyman, and to convey what is fundamentally the same notion by specifying the confidence coefficient associated with an interval. Thus the confidence coefficient may be regarded as a particular value of the fiducial probability selected to form the basis of the calculation, to be employed in repeated experience, of the confidence interval*.

The method of solution of the problem may be illustrated with the help of Fig. 1, in which $n = 10$; p and x have been taken as coordinate axes, so that p may lie between 0 and 1, while x may assume any of the integral values 0, 1, ... 10. In our experience with samples of 10 individuals, no point (x, p) can lie outside the square of the diagram. For a given value of p, the chance of occurrence of different values of x will be given by the terms of the binomial expansion $(q+p)^{10}$. Let (a) $S(p, n;\ 0 \ldots x)$, and (b) $S(p, n;\ x \ldots n)$, denote the sum of (a) the 1st $x+1$, and (b) the last $n-x+1$ terms. Then while it will not in general be possible to choose values of x_1 and x_2 so that both $S(p, n;\ 0 \ldots x_1)$ and $S(p, n;\ x_2 \ldots n)$ equal exactly some selected value, say ·025, it will be possible to choose x_1 and x_2 so that

$$S(p, n;\ 0 \ldots x_1) \leqslant \cdot 025 < S(p, n;\ 0 \ldots x_1 + 1) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4),$$

$$S(p, n;\ x_2 \ldots n) \leqslant \cdot 025 < S(p, n;\ (x_2 - 1) \ldots n) \quad \ldots\ldots\ldots\ldots\ldots\ldots (5).$$

The position is illustrated diagrammatically below:

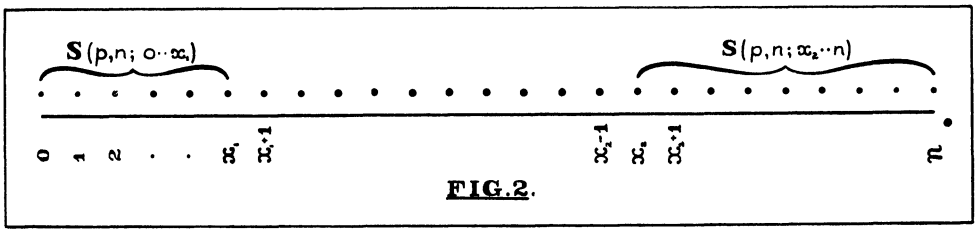

FIG. 2.

If it is supposed that such values are determined for x_1 and x_2 throughout the whole range, $p = 0$ to 1, we shall obtain two series of stepped lines running across the diagram as shown in Fig. 1, all points on which satisfy conditions (4) and (5) respectively. It follows that in the long run of our statistical experience from whatever populations random samples of 10 are drawn, we may expect at least 95 °/₀ of the points (x, p) will lie inside the lozenge shaped belt, not more than

* J. Neyman, "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection," *Journal of Royal Statistical Society*, XCVII. pp. 558—606, 1934.

$2\frac{1}{2}$ % on or above the upper boundary and not more than $2\frac{1}{2}$ % on or below the lower boundary. If then as a general rule, when x alone is known these boundaries are used to determine points (x, p_1) and (x, p_2), we may have confidence that we shall be correct in the estimate $p_1 < p < p_2$ in about 95 % of cases. If greater confidence is desired, we may determine wider limits leading to a higher value of the expected percentage accuracy, e.g. 99 %. In the diagram, values of p_1 and p_2 corresponding to $x = 2$ are shown.

This plan has been carried out below with the following modifications adopted for practical convenience:

(1) The charts prepared are entered with p and $\theta = x/n$, so that $0 \leqslant \theta \leqslant 1$, and boundaries for a number of values of n can be drawn on the same chart.

(2) Instead of the stepped boundaries, curves have been drawn as in Fig. 1 passing through the inner "corner" points, i.e. the points $(x_1 p)$ and $(x_2 p)$ for which p is such that $S(p, n; 0 \ldots x_1)$ and $S(p, n; x_2 \ldots n)$ are exactly equal to the desired chance (in the cases chosen, these chances are ·025 and ·005). These curves are more convenient than the stepped lines for interpolation for intermediate values of n. Since no possible point (x, p) can fall inside the area between a curve and the steps, no error is involved in using the curves.

(3) While the "corner" points could have been calculated precisely and the curves drawn through them, it was considered sufficiently accurate for the purpose to obtain the curves by an approximate method of interpolation described below.

Before describing the charts and illustrating their use, it may be well to make clear the sense in which this method of estimation in terms of a confidence or fiducial interval does not depend on any *a priori* knowledge regarding possible values of p. Consider the following situation. Suppose that in the course of our experience samples of 30 are continually drawn, and that although we are not aware of the fact, these are taken from populations in which p has three different values only, namely $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$. Further that the proportions of times these three cases are met with are as $\frac{8}{10} : \frac{1}{10} : \frac{1}{10}$ respectively.

The expectation, on a basis of 10,000 draws, is shown in Fig. 3, in which the axes of p and x have been reversed for convenience. For example, for $p = \frac{1}{3}$, $x = 12$, the expectation is 881, while for $p = \frac{2}{3}$, $x = 28$, it is 1. The chart of Fig. 4 described below will provide for each of the 31 possible values of x the limits for the confidence interval, with coefficient ·95, for p. Thus when $x = 15$, we find $p_1 = ·31$, $p_2 = ·69$. Taken over the whole experience, these intervals include the true population value of p in 9676 out of the 10,000 cases, and in the remaining 324 do not, that is to say we are wrong in less than 5 % of cases. This is the risk of error that we have accepted, and it is quite independent of the particular set of three values of p introduced, or the relative frequency with which they are encountered in our experience.

It will be noticed, however, from the figures in the margin, that the percentage of wrong judgments differs according to the value of x, from 100 to 0. We cannot

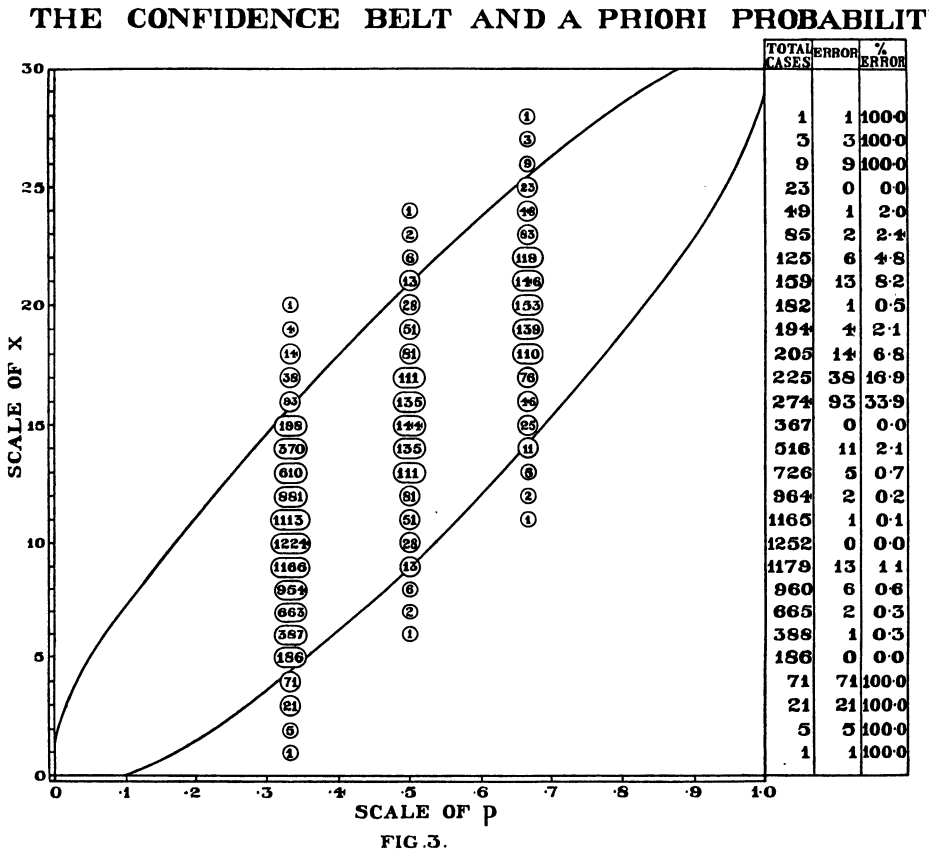

FIG. 3.

therefore say that for any specified value of x the probability that the confidence interval will include p is ·95 or more. The probability must be associated with the whole belt, that is to say with the result of the continued application of a method of procedure to all values of x met with in our statistical experience.

Indeed it will be clear that *if* we had information *a priori* regarding the values of p likely to be met in our experience, and if this information could be expressed in precise numerical form, it would be possible to shift the confidence belt and so narrow the limits of uncertainty while retaining the same risk of error. For instance, if we knew that $\frac{1}{3} \leqslant p \leqslant \frac{2}{3}$, we should certainly cut off the two points of the lozenge by lines at $p = \frac{1}{3}$ and $p = \frac{2}{3}$.

In practice, however, it is rare

(1) for the *a priori* information to be expressed in exact form,

(2) even when it appears so expressible, for the working statistician to have time to calculate suitable modification for the limits.

Under general conditions, therefore, the statistician will usually be satisfied with limits which are "safe" in the sense that they give an expectation of long run

accuracy which is precisely known*, and thus avoid the uncertain risk of error involved in an attempt to introduce *a priori* information.

(2) *Calculation and use of the charts.*

The following method was employed in obtaining points from which to draw the curves in Figs. 4 and 5.

Samples with $n = 10, 15, 20, 30$.

Use was made of the tables giving the continued sum of the binomial terms, published in one of the Medical Research Council's Reports†. From these tables it is possible to find the sum of any number of binomial terms for $p = \cdot025, \cdot05, \cdot075, \cdot10, \cdot15, \cdot20, \ldots, \cdot85, \cdot90, \cdot925, \cdot95, \cdot975$. It will happen only rarely that for these values of p, $S(p, n; 0 \ldots x_1)$ or $S(p, n; x_2 \ldots n)$ approach the desired values of ·025 or ·005, i.e. that we can obtain directly the inside "corners" of the steps of Fig. 1. For the purpose of the charts, however, it was considered that sufficient accuracy would be obtained by interpolation for x in the tables. Take for example the case of $n = 20$ and consider the sums of the binomial terms for $p = 0{\cdot}45$ given below. At what points should the two curves associated with $n = 20$ cut the lines

x	$S(\cdot45, 20; 0 \ldots x)$	x	$S(\cdot45, 20; 0 \ldots x)$
0	·0000	8	·4143
1	·0001	9	·5914
2	·0009	10	·7507
3	·0049	11	·8692
4	·0189	12	·9420
5	·0553	13	·9786
6	·1299	14	·9936
7	·2520	15	·9985

$p = \cdot45$ in the charts? The point $x = 3$ $(x/n = \cdot150)$ is approximately a "corner" point, since the sum of the first 4 terms equals almost exactly ·005, but the other points must be obtained by interpolation. Thus we argue:

(*a*) *The lower* ·025 *point.* The sum of the terms 0...4 is ·0189 (< ·025), and the sum of the terms 0...5 is ·0553 (> ·025); a linear interpolation‡ gives for x, 4·17 $(x/n = \cdot208)$.

(*b*) *The upper* ·025 *point.* The sum of the terms 13...20 is $1 - \cdot9420 = \cdot0580$, and the sum 14...20 is $1 - \cdot9786 = \cdot0214$. Take $x = 13{\cdot}90$ $(x/n = \cdot695)$.

(*c*) *The upper* ·005 *point.* The sum of the terms 15...20 is $1 - \cdot9936 = \cdot0064$, and the sum 16...20 is $1 - \cdot9985 = \cdot0015$. Take $x = 15{\cdot}29$ $(x/n = \cdot764)$.

* This is not strictly true of course, since only an upper limit to the error is known, owing to the fact that x can assume discrete values only. As n increases, however, the true risk will rapidly approach the limiting value. In cases where the coefficient, x, is a continuous variable such as a sample mean or standard deviation, this difficulty does not arise.

† *Reports on Biological Standards*, "II. Toxicity Tests for Novarsenobenzene," by Durham, Gaddum and Marchal.

‡ This form of interpolation, if crude, appeared adequate for the curve drawing.

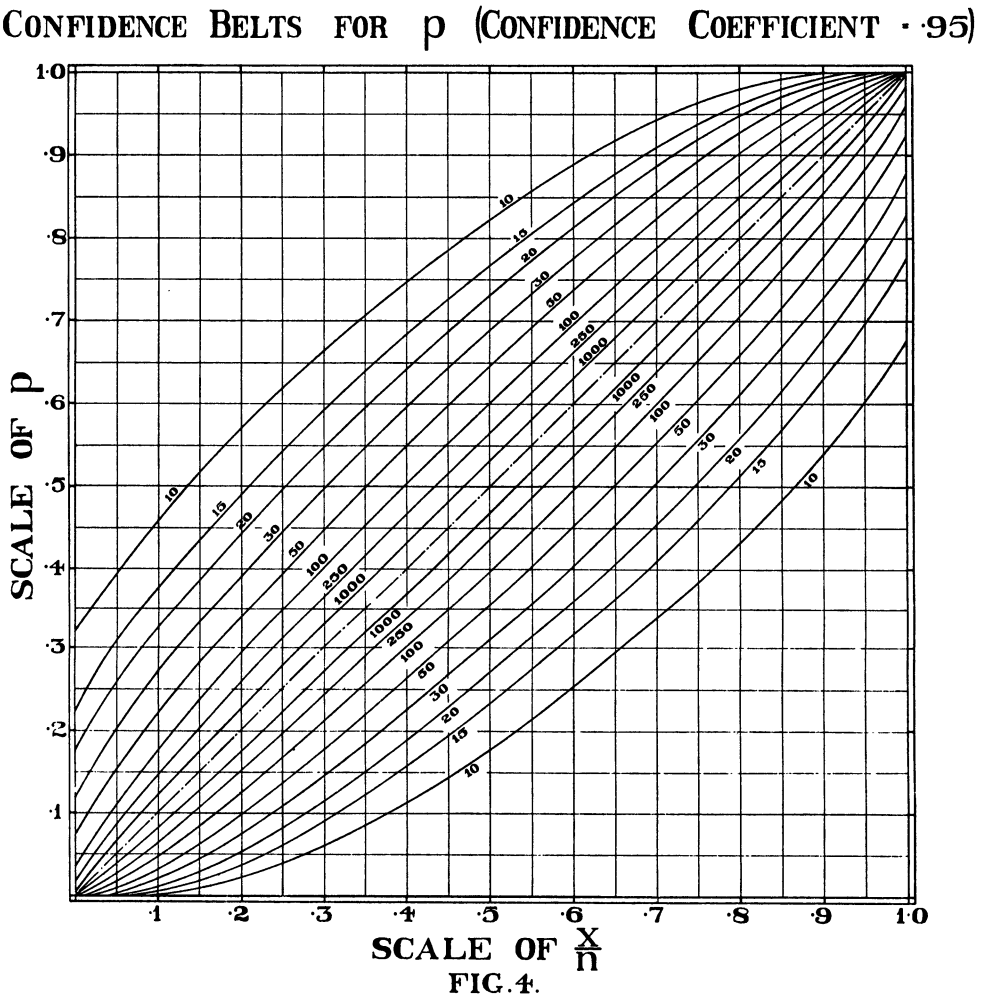

FIG. 4.

Consequently in Fig. 4 the curves marked $n = 20$ cut the line $p = \cdot 45$ at $x/n = \cdot 208$ and $\cdot 695$, and in Fig. 5 at $x/n = \cdot 150$ and $\cdot 764$.

Fresh calculations of binomial terms were made for $n = 50$, 100 and 250, while the limits were obtained from the normal curve in the case $n = 1000$.

It will be noted that the curves cut the axis $x/n = 0$ at points at some distance from $p = 0$ when n is small. The points of intersection correspond in the two diagrams to those values of p for which the first term of the binomial $q^n = (1-p)^n$ equals ·025 and ·005 respectively. On the other side, the end points on the axis $x/n = 1$ correspond to values of p for which the last term, p^n, equals ·025 and ·005.

The charts have been prepared to give rapid answers in problems such as the following:

(1) A sample has been drawn (n and x known), to obtain the confidence or fiducial interval for p.

CONFIDENCE BELTS FOR p (CONFIDENCE COEFFICIENT ··99)

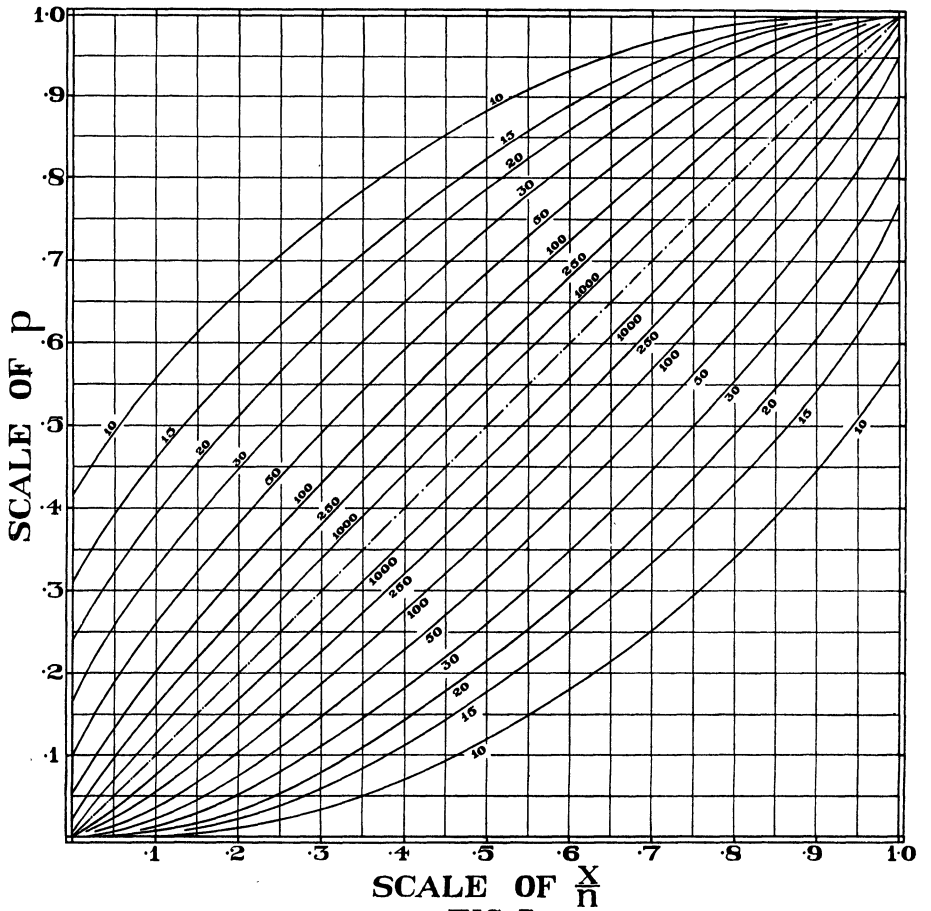

FIG. 5.

Example A. The toxicity of a drug may be measured by the proportion, p, of mice in a standard laboratory population that will die after injection with a dose of given strength. Out of a sample of 30 mice randomly selected from the population, 8 die after injection; within what limits may we expect that p lies? Turning to Fig. 4, and taking $n = 30$, $x/n = 8/30 = \cdot 267$, it will be seen that we may say that $\cdot 12 < p < \cdot 46$, if we are prepared to accept a risk of error of not more than 1 in 20. To obtain greater confidence in prediction (risk of error 1 in 100) we must turn to Fig. 5 and obtain $\cdot 09 < p < \cdot 52$.

(2) To plan in advance the size of sample necessary to provide a desired degree of accuracy in estimation.

Example B. In a manufacturing process a crude index of quality, P, has been the percentage of articles which pass a certain test. This index has fluctuated in

the past round $P = 60$, but it is proposed to make an intensive effort to improve quality (which will mean the raising of this percentage) by tightening the control of manufacture. Improvement is to be judged by studying the changes in the proportion of articles (x/n) passing the test in a random sample of n articles. How large would n need to be to obtain from the sample an estimate of P, with a range of uncertainty of not more than 5 ?

At the start, the value of $p = P/100$ in the material sampled is not more than ·60, and we wish to determine n so that the confidence belt will be of breadth about ·05. On the assumption that a confidence coefficient of ·95 is adequate, we may use Fig. 4. It will be seen that for x/n having values between ·6 and ·8, n must be more than 1000 for the interval $p_2 - p_1$ to be as small as ·05*. In many cases the testing of so large a sample would be quite out of the question, and this result points to the fact that an index of this type is not an efficient measure of quality. Much more information of changes could probably be drawn from a smaller sample, if the index could be based on the mean value of some measured character determined for each article of the sample.

(3) To determine the limits of sampling variation that may be expected in x when p is known, and so determine the size of sample needed.

Example C. There are two alternative hypotheses regarding the chance of an individual in a certain population bearing a given character; the alternatives are that $p = \frac{1}{4}$ or $p = \frac{1}{2}$. Such might be the case in some genetic investigation. How large a sample must be planned to make it practically certain that we can discriminate between the two hypotheses ?

In this case we are concerned with the sampling variation of x for $p = \frac{1}{4}$ and $p = \frac{1}{2}$, and n should be chosen so large that there is no "overlap" of any consequence between the two distributions. Suppose we choose n so that the upper ·005 point of the x distribution for $p = \frac{1}{4}$, as judged from the curves of Fig. 5, corresponds to the lower ·005 point of the distribution for $p = \frac{1}{2}$. This will occur when n is slightly over 100, say 110†.

* Since for large values of n the upper and lower bounds of the confidence belt are very nearly parallel lines making an angle of 45° with the axes, and the binomial may be represented by a normal curve, the breadth of the belt is approximately $4\sqrt{p(1-p)/n}$, which if equated to ·05 gives $n = 1600$ for $p = ·60$ and $n = 1000$ for $p = ·80$.

† [It is interesting to consider what the solution would be, if the two binomials, $(p_1 + q_1)^n$ and $(p_2 + q_2)^n$, were replaced by normal curves. The means of these curves will be np_1 and np_2 $(p_2 > p_1)$ while their standard deviations will be $\sigma_1 = \sqrt{np_1q_1}$ and $\sigma_2 = \sqrt{np_2q_2}$. Let l represent the overlap and x_1 represent the distance from mean to overlap in first curve and x_2 represent distance from overlap to mean in second curve. Accordingly

$$n(p_2 - p_1) = x_1 + x_2 \quad \text{.....(i).}$$

If l be the overlap, and $\frac{1}{2}(1 - \alpha_1)$ be the area cut off from the first curve and $\frac{1}{2}(1 - \alpha_2)$ from the second curve, it will be reasonable to take

$$\tfrac{1}{2}(1 - \alpha_1) = \tfrac{1}{2}(1 - \alpha_2) = \tfrac{1}{2}l.$$

Thus x_1/σ_1 and x_2/σ_2 must be obtained from the tables of the normal probability integral with $\frac{1}{2}(1 + \alpha) = 1 - \frac{1}{2}l$, or say they have the value ξ.

If we were prepared to accept a greater risk of an inconclusive result, which we might well be prepared to do if the sample could be readily increased in size in a doubtful case, then we might choose n so that the upper and lower ·025 points of the x distributions correspond. Turning to Fig. 4, it is found that this occurs when n is about 65.

It follows from (i) that

$$n(p_2 - p_1) = \xi(\sqrt{np_1q_1} + \sqrt{np_2q_2})$$

or

$$\sqrt{n} = \frac{\xi(\sqrt{p_1q_1} + \sqrt{p_2q_2})}{p_2 - p_1} \quad \text{.....(ii).}$$

In the case in the text $p_1 = \frac{1}{4}$, $p_2 = \frac{1}{2}$ and $l = \cdot 005$, $1 - \frac{1}{2}l = \cdot 9975$, which corresponds to $\xi = 2 \cdot 81$ nearly. Thus $\sqrt{n} = 2 \cdot 81\,(\sqrt{3} + 2) = 2 \cdot 81 \times 3 \cdot 7205 = 10 \cdot 455$ and $n = 109 \cdot 3$, according well with the value in the text. If we desire to alter the overlap, the first factor ξ only is changed in (ii). If $l = \cdot 025$, then $\frac{1}{2}(1 + \alpha) = \cdot 9875$ and $\xi = 2 \cdot 2416$ $\sqrt{n} = 2 \cdot 2416 \times 3 \cdot 7205 = 8 \cdot 34$ and $n = 69 \cdot 6$, or 70 as against 65 of text above. ED.]