ON CONFIDENCE LIMITS AND SUFFICIENCY, WITH PARTICULAR REFERENCE TO PARAMETERS OF LOCATION

By B. L. Welch

1. **Introduction.** The solution of the problem of estimating an interval in which a population parameter should lie, by means of what is now often termed the fiducial type of argument, dates back to the early writers on the theory of errors. However, owing to their lack of "Student's" z distribution, their statements were usually only of an approximate character, and, furthermore, the logical distinction between the fiducial method and the method of inverse probability was never clearly drawn, before R. A. Fisher discussed the subject. It is of interest to note how far "Student" himself went in this matter. In describing the tables which he gave in his original paper he says:[1]

"The tables give the probability that the value of the mean, measured from the mean of the population, in terms of the standard deviation of the sample, will lie between $-\infty$ and z. Thus, to take the tables for samples of six, *the probability of the mean of the population lying between* $-\infty$ *and once the standard deviation of the sample is* 0.9622 or the odds are about 24 to 1 that the mean of the population lies between these limits. The probability is therefore 0.0378 that it is greater than once the standard deviation, and 0.0756 that it lies outside $\pm$ 1.0 times the standard deviation."

It should be noted that "Student's" z is $(\bar{x} - \theta)/s$ where θ is the true population mean. His tables tell us that for $n = 6$, $P(z < 1)$[2] is equal to 0.9622. Owing to the symmetry of the z distribution this is equivalent to saying that $P(z > -1)$ is 0.9622, i.e.

$$P\left\{\frac{\bar{x} - \theta}{s} > -1\right\} = 0.9622.$$

This may be transposed to read

$$P\{\theta < \bar{x} + s\} = 0.9622 \tag{1}$$

which is the statement I have italicized in the above extract, it being there understood that the mean of the population is being measured from the mean of the sample. "Student" therefore makes here what is now called a fiducial statement. In the next sentence he, in effect, attaches a probability to an interval estimate for the population mean. In doing this "Student" was not conscious of introducing any new principle, nor does he apply the method consistently

[1] "Student" (1908). "The Probable Error of a Mean." *Biometrika* VI, p. 20.

[2] P is used to denote the probability of the truth of the relation in the bracket following.

to other problems of estimation. For instance, in discussing the estimation of the correlation coefficient ρ about the same time, he formulates the problem in terms of inverse probability, although he was fully aware of the difficulties involved in postulating an *a priori* distribution for ρ.

In discussing the problem of interval estimation more generally, I shall adopt some of the terminology used by J. Neyman.[3] The sample observations $x_1, x_2, \cdots x_n$ will be noted collectively by E (standing for the "event" point when the observations are represented as coördinates in a space of n dimensions). Then if θ is an unknown parameter, α a fixed probability, and $F(E, \theta, \alpha)$ a function such that

$$P\{F(E, \theta, \alpha) > 0\} = \alpha \tag{2}$$

we may obtain an interval estimate for θ as follows. Let $\delta(E, \alpha)$ denote the set of values of θ such that for any θ in the set we have $F(E, \theta, \alpha) > 0$. Then if we use the notation $\{\delta(E, \alpha) \subset \theta\}$ to indicate that the set $\delta(E, \alpha)$ contains or "covers" the true parameter θ we shall be able to rewrite (2)

$$P\{\delta(E, \alpha) \subset \theta\} = \alpha. \tag{3}$$

We can then adopt the following rule to obtain an interval estimate for θ: (a) calculate from the sample the set $\delta(E, \alpha)$, (b) make the statement that $\delta(E, \alpha)$ covers θ. In adopting this rule we shall be right in the proportion α of cases.

There are, in general, an infinite number of ways in which we can start with a statement of the type (2) to reach the statement of type (3). Neyman has discussed methods of making the best choice between such statements. His approach to this problem may be illustrated by the following example.

Suppose we have a random sample of n from a normal population with standard deviation σ and let

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n - 1)},$$

and w = range = largest x − smallest x.
Then we can find a constant b_α such that

$$P\left\{\frac{s}{\sigma} > b_\alpha\right\} = \alpha \tag{4}$$

and, turning this round, we obtain

$$P\left\{\sigma < \frac{s}{b_\alpha}\right\} = \alpha. \tag{5}$$

This means that, if we choose $\alpha = .99$ (say), then we can say that σ is less than $s/b_{.99}$ and in 99% of cases we shall be correct in this statement.

---

[3] J. Neyman (1937). "Outline of a theory of statistical estimation based on the classical theory of probability." *Phil. Trans. Roy. Soc.* A 236, pp. 333–380.

Now similarly we can find c_α such that

$$P\left(\frac{w}{\sigma} > c_\alpha\right) = \alpha \tag{6}$$

and reversing this

$$P\left(\sigma < \frac{w}{c_\alpha}\right) = \alpha. \tag{7}$$

This statement is not inconsistent with (5). It means that, if we choose to base our rule of estimation *always* on the range, then in 99% of cases we shall be correct in saying that $\sigma < w/c_{.99}$. On the other hand, (5) relates to the consequences of applying *always* a rule of estimation based on the standard deviation of the sample. Both (5) and (7) are in themselves true statements, but we must decide which of them is the better one to use. In certain circumstances speed of calculation may be the determining factor, in which case (7) may be preferable, but here we shall assume that the time spent on calculation is not important.

In making the statement that σ is less than some upper limit which is a function of the sample observations, we shall, in general, prefer that this upper limit be placed as low as possible consistent with the chosen confidence coefficient α. We find, however, that it is not possible to say that, whatever the sample obtained, s/b_α will be less than w/c_α or *vice versa*. We must, therefore, approach the problem from another angle. If σ' is a value greater than the true standard deviation σ we can theoretically evaluate the *probability* that $\sigma' < s/b_\alpha$, and similarly the probability that $\sigma' < w/c_\alpha$. We may now express our general desire to place the upper confidence limit for σ as low as possible in a more concrete form. We may ask that the probability that σ' is less than this limit should be as small as possible. We find in the present problem that, whatever $\sigma' > \sigma$, we should include σ' in the interval from 0 to s/b_α less frequently than we should in an interval based on any other statistic. This constitutes an argument for using s rather than any other statistic such as w.

In general, Neyman makes all problems of choosing between alternative procedures of interval estimation depend on the probability that the intervals include values of the parameter different from the true value, as well as on the probability of them containing the true value. This principle of choice does, I think, appear reasonable, although its application is not, of course, so straightforward when statistics with properties of sufficiency similar to those of s do not exist. It is then necessary to introduce other conditions into the formulation of the problem. I intend to discuss elsewhere ways in which this has been done.

To summarize, we may say: (a) we can make many true statements of the type (3); and (b) if we can agree on certain further properties which these statements should possess, we can choose which is the best statement of this type to adopt as our general rule for interval estimation. There are certain differences

between this approach and that of R. A. Fisher, whose attitude is expressed clearly in his contribution to the discussion following Neyman's paper[4] "On the two different aspects of the representative method." Fisher says there that: "In particular he would apply the fiducial method, or rather would claim unique validity for its results, only in those cases for which the problem of estimation proper had been completely solved, i.e. either when there existed a statistic of the kind called *sufficient*, which in itself contained the whole of the information supplied by the data, or when, though there was no sufficient statistic, yet the whole of the information could be utilized in the form of *ancillary* information." Thus it appears that when sufficient statistics do not exist, excepting in those further cases where Fisher claims that the problem of estimation has been completely solved, he would definitely discourage the use of the fiducial argument at all. Neyman, on the other hand, would allow the attempt to obtain interval estimates on the lines described above. Where sufficient statistics do exist, the two approaches do not lead to any final disagreement. Neyman, using results obtained in the Neyman-Pearson theory of testing hypotheses, is led to criteria depending in a particular way on the joint probability law of the sample, and these criteria are seen to involve the sample values only through statistics which have been defined as sufficient. One may regard this fact in two ways: (a) one may say that because a certain line of approach, which seems intuitionally sound, leads to the use of statistics which have been defined as sufficient, therefore this definition of sufficiency is a good one, or (b) one may say that the definition of sufficient statistics is fundamental, and that any method of approach which leads to their use has thereby obtained some extra support.

There remains the case alluded to above, where the joint probability law of the sample does not depend on the unknown parameter θ by way of one statistic only, but where nevertheless it has been said that the problem of estimation has been completely solved. This case will be discussed in the next section.

2. Interval Estimates of Location. R. A. Fisher has given, as a particular example, a case where the unknown parameter is one of location, so that we can write

$$p(x \mid \theta) = \phi(x - \theta).$$

Now if we have a sample of n from this distribution, the $(n - 1)$ differences between successive observations when arranged in order of magnitude will have a joint distribution independent of θ. Hence if we denote the sample by E, and the $(n - 1)$ differences jointly by C, we have

$$p(E \mid \theta) = p(T \mid C, \theta)p(C) \tag{8}$$

where T is some statistic, such as the mean or median, whose distribution does depend on θ and may hence be taken as an estimate of θ. We may therefore

[4] J. Neyman (1934). *J. R. Statist. Soc.* 97, p. 617.

read (8) as follows: the joint probability law of the sample is equal to the probability law of the estimate in samples of the same configuration, C, multiplied by the probability of the configuration, the latter not depending on the unknown θ. From this it has been deduced that all the information respecting θ provided by the sample is given by referring T to the distribution $p(T \mid C, \theta)$. Fisher,[5] for instance, says that "in interpreting our estimate (we) may take as its sampling distribution that appropriate to only those samples which have the actual configuration observed." Later in the same context he remarks that in general, when θ is a parameter of any type whatever, and not necessarily one of location or scaling, if something can be found "corresponding with the configuration of the sample in the simple case discussed above, . . . one of the primary problems of uncertain inference will have reached its complete solution. If not, there must remain some further puzzles to unravel."

It is clear, therefore, that more has been claimed for this method than that it is *practically* useful, or that it yields the best results possible in *large* samples, or that it yields results *highly approximating* the best possible in small samples. There is an emphasis here on completeness that leads one to suppose that all problems of estimation and testing hypotheses may be answered to the best advantage by considering only the distribution of an estimate in samples of the same configuration, the estimate thus attaining properties analogous to those of a sufficient statistic. That this supposition is not true may be seen by considering the following simple example. This example concerns the simplest situation that one deals with in the theory of testing statistical hypotheses. Its relevance to the problem of interval estimation will, however, not be difficult to see.

Suppose that we have a sample from a population involving only a parameter of location θ, and that we wish to test whether θ is equal to θ_0 (say), and that besides θ_0 there is only one value θ_1 (say) which it is possible for θ to take. Suppose we require to set up a statistical test which will reject the hypothesis $\theta = \theta_0$, in only a small proportion ϵ of cases, when it is true. Many such tests are possible, and it is natural to choose from them that test which will lead most frequently to the rejection of the hypothesis that $\theta = \theta_0$ when the single alternative $\theta = \theta_1$ is true. Neyman and Pearson[6] have shown that the best test from this viewpoint is provided by the criterion

$$J = \frac{p(E \mid \theta_1)}{p(E \mid \theta_0)}. \tag{9}$$

This criterion must be referred to its distribution in *all* samples when $\theta = \theta_0$. We must therefore choose a constant J_ϵ such that

$$P(J > J_\epsilon \mid \theta = \theta_0) = \epsilon \tag{10}$$

[5] Fisher, R. A. (1936). "Uncertain Inference." *Proc. Amer. Acad. Arts and Sciences*, 71, No. 4, p. 257.

[6] J. Neyman and E. S. Pearson (1932). "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Roy. Soc.* A 231, p. 300.

and reject the hypothesis that $\theta = \theta_0$ when $J > J_\epsilon$. This is known to be the best test in these circumstances, and we may demand that any other procedure which claims to use the data exhaustively should be equivalent to it. Now if we decide to use only the distribution of the statistic T in samples of the same configuration, we are led to take as the most powerful test based on $T \mid C$ one which would reject the hypothesis that $\theta = \theta_0$ when the ratio of $p(T \mid C, \theta_1)$ to $p(T \mid C, \theta_0)$ exceeds a certain value. Now by (8) this ratio is exactly the criterion J of (9) above. There is, however, this difference, that J has now to be referred to its distribution in samples with the *same configuration* C as that observed. We shall therefore have to choose $J_\epsilon(C)$ such that

$$P\{J > J_\epsilon(C) \mid C, \theta\} = \epsilon. \tag{11}$$

A test, then, which rejects the hypothesis that $\theta = \theta_0$ when $J > J_\epsilon(C)$ will be such that it is the most powerful possible with respect to the alternative $\theta = \theta_1$, based on samples with the same configuration. However, in actual sampling from a population, we derive samples with all configurations, and the real power of the test will therefore be measured by

$$P\{J > J_\epsilon(C) \mid \theta_1\} = \int P\{J > J_\epsilon(C) \mid C, \theta_1\} p(C)\, dC. \tag{12}$$

This quantity cannot be greater, and will in general be less, than the power[7] of the other test, viz. $P(J > J_\epsilon \mid \theta_1)$. (If $J_\epsilon(C)$ is the same for all C, and therefore equal to J_ϵ, the powers will be equal. This will be the case when there is a sufficient statistic for θ.) We must therefore conclude that, in relation to this simple problem at least, a method which takes account only of distributions in samples with the same configuration will not use the data to the best advantage.

Of course the type of problem to be solved is usually not so straightforward as the present one. There will usually be more than one value of θ alternative to θ_0, and no uniformly most powerful test will, in general, exist. It is legitimate, however, to consider the above example, because any procedure claiming properties of sufficiency should be able to deal with it in the best possible way.

An example may make the above points clearer, and will show their relevance to the problem of interval estimation. Consider a rectangular distribution with mean θ, and range from $(\theta - \frac{1}{2})$ to $(\theta + \frac{1}{2})$. Let x_1 and x_2 be a sample of 2 from this population, and suppose we require confidence limits for θ such that the chance of them enclosing θ is α.

If we represent x_1 and x_2 as coördinates of a point with respect to rectangular axes, the joint probability distribution is constant over a square centered at the point (θ, θ). This is shown by $ABCD$ in Fig. 1. We have

$$p(x_1, x_2)dx_1dx_2 = dx_1dx_2 \begin{cases} \theta - \frac{1}{2} < x_1 < \theta + \frac{1}{2} \\ \theta - \frac{1}{2} < x_2 < \theta + \frac{1}{2}. \end{cases} \tag{13}$$

[7] Power is used throughout in the Neyman-Pearson sense, i.e. to denote the chance of a test rejecting a hypothesis when a given alternative is actually true.

If we write $z_1 = \frac{1}{2}(x_1 + x_2)$; $z_2 = \frac{1}{2}(x_1 - x_2)$, z_2 will represent the configuration of the sample, and z_1 may be taken as the estimate, T, of θ in our discussion above. We can then show that

$$p(z_1, z_2)dz_1dz_2 = 2dz_1dz_2, \tag{14}$$

$$p(z_2)dz_2 = 2\{1 - 2\,|\,z_2\,|\,\}dz_2 \cdots -\tfrac{1}{2} < z_2 < \tfrac{1}{2}, \tag{15}$$

and

$$p(z_1\,|\,z_2)\,dz_1 = \frac{dz_1}{1-2\,|\,z_2\,|} \cdots \theta - \tfrac{1}{2} + |\,z_2\,| < z_1 < \theta + \tfrac{1}{2} - |\,z_2\,|. \tag{16}$$

That these are the correct limits for z_1 and z_2 may be seen by reference to Fig. 1, noting that z_1 and z_2 are constant along lines parallel to the respective diagonals BD and AC of the square.

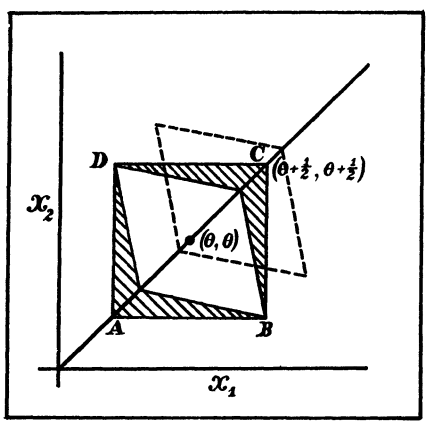

FIG. 1

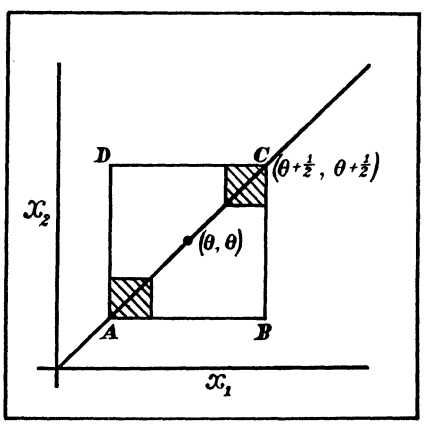

FIG. 2

First let us confine ourselves to samples with the same configuration z_2. Then, from (16), we can say that

$$P\{\theta - \alpha(\tfrac{1}{2} - |\,z_2\,|) < z_1 < \theta + \alpha(\tfrac{1}{2} - |\,z_2\,|)\} = \alpha. \tag{17}$$

This statement is true for given z_2, and will be *a fortiori* true when this restriction is removed. It is equivalent to saying that the chance of a point falling into the shaded area in Fig. 1 is $(1 - \alpha)$, where α denotes the proportion of the diagonal AC lying in the non-shaded area.[8] Confidence limits for θ are then obtained by transposing (17), giving

$$P\{z_1 - \alpha(\tfrac{1}{2} - |\,z_2\,|) < \theta < z_1 + \alpha(\tfrac{1}{2} - |\,z_2\,|)\} = \alpha. \tag{18}$$

[8] We are assuming that confidence limits are required such that the chance is $\left(\frac{1}{2} - \frac{\alpha}{2}\right)$ of θ being above the upper limit, and $\left(\frac{1}{2} - \frac{\alpha}{2}\right)$ of it being below the lower limit.

That this is not the best way of constructing confidence limits is seen as follows. Let us denote the lesser of x_1 and x_2 by x_L, and the greater by x_G. Then if we consider the possible values of x_L and x_G which will satisfy simultaneously the inequalities

$$(19) \qquad \begin{cases} \theta - \tfrac{1}{2} < x_L < \theta + \tfrac{1}{2} - \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \\ \theta - \tfrac{1}{2} + \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} < x_G < \theta + \tfrac{1}{2} \end{cases}$$

we see that they lie in the non-shaded area of the square $ABCD$ in Fig. 2 where the sides of the shaded squares are $\sqrt{\frac{1}{2} - \frac{\alpha}{2}}$. The chance of the inequalities holding simultaneously is therefore α. Further we see that these inequalities can be transposed to read

$$(20) \qquad \begin{cases} x_G - \tfrac{1}{2} < \theta < x_L + \tfrac{1}{2} \quad \text{when } (x_G - x_L) > \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \\ x_L - \tfrac{1}{2} + \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} < \theta < x_G + \tfrac{1}{2} - \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \\ \qquad\qquad \text{when } (x_G - x_L) < \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \end{cases}$$

and therefore we can take these to define our confidence limits for θ.

The intervals defined by the confidence limits in (18) and (20) are equivalent in the sense that each covers the true value of θ in a proportion α of cases. To decide which is the better rule of interval estimation we shall follow Neyman, and consider how often the intervals cover values other than the true θ. In particular let $(\theta + \Delta)$ be any other value, and consider the expressions P_1 and P_2 where

$$(21) \qquad P_1 = P\{z_1 - \alpha(\tfrac{1}{2} - |z_2|) < (\theta + \Delta) < z_1 + \alpha(\tfrac{1}{2} - |z_2|)$$

and P_2 is the probability that one or another of the following inequalities holds

$$(22) \qquad \begin{cases} x_G - \tfrac{1}{2} < (\theta + \Delta) < x_L + \tfrac{1}{2} \quad \text{when } (x_G - x_L) > \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \\ x_L - \tfrac{1}{2} + \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} < (\theta + \Delta) < x_G + \tfrac{1}{2} - \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \\ \qquad\qquad \text{when } (x_G - x_L) < \sqrt{\tfrac{1}{2} - \tfrac{\alpha}{2}} \end{cases}$$

Now (21) can be written

$$(23) \qquad P_1 = P\{(\theta + \Delta) - \alpha(\tfrac{1}{2} - |z_2|) < z_1 < (\theta + \Delta) + \alpha(\tfrac{1}{2} - |z_2|)\}.$$

Referring to Fig. 1 we see that we have to evaluate the chance of the sample falling into a lozenge-shaped area like the unshaded area in $ABCD$, but moved bodily along the diagonal AC to such a position as is indicated by the dotted lines. Difficulties are introduced by the discontinuities, but we can show that for Δ positive

$$(24)\quad \begin{cases} P_1 = \alpha \quad \text{when } \Delta = 0 \\ P_1 = \alpha - \dfrac{4\alpha\Delta^2}{1-\alpha^2} \cdots 0 \le \Delta \le \left(\dfrac{1}{2} - \dfrac{\alpha}{2}\right) \\ P_1 = \left(\dfrac{1}{2} + \dfrac{\alpha}{2}\right) - 2\Delta + \dfrac{2\Delta^2}{1+\alpha} \cdots \left(\dfrac{1}{2} - \dfrac{\alpha}{2}\right) \le \Delta \le \left(\dfrac{1}{2} + \dfrac{\alpha}{2}\right) \\ P_1 = 0 \qquad \Delta \ge \left(\dfrac{1}{2} + \dfrac{\alpha}{2}\right) \end{cases}$$

with similar expressions for Δ negative. The graph of P_1 against Δ is shown in Fig. 3, α for convenience being taken $= 0.92$. From it we can read off the probability of the confidence interval covering $(\theta + \Delta)$, where θ is the true value of the parameter.

Similar calculations may be made for P_2. Without going into details, it is seen that

$$(25)\quad \begin{cases} P_2 = \alpha & \text{when } \Delta = 0 \\ P_2 = \alpha - 2\Delta\left(1 - \sqrt{\dfrac{1}{2} - \dfrac{\alpha}{2}}\right) & 0 \le \Delta \le \sqrt{\dfrac{1}{2} - \dfrac{\alpha}{2}} \\ P_2 = \left(\dfrac{1}{2} + \dfrac{\alpha}{2}\right) - 2\Delta + \Delta^2 & \sqrt{\dfrac{1}{2} - \dfrac{\alpha}{2}} \le \Delta \le 1 - \sqrt{\dfrac{1}{2} - \dfrac{\alpha}{2}} \\ P_2 = 0 & \Delta \ge 1 - \sqrt{\dfrac{1}{2} - \dfrac{\alpha}{2}} \end{cases}$$

P_2 is plotted against Δ in Fig. 3. It is seen that, whatever value of Δ we take, the chance of $(\theta + \Delta)$ being included in the confidence interval, is less for the second method of estimation than it is for the method based on the distribution of $z_1 \mid z_2$.[9] This circumstance would, I think, contradict the view that the latter method was deriving the utmost from the sample. Whether the method is still a good one, though not necessarily the best, is not a question at issue in the present paper. The curves in Fig. 3 are very close together, and we are led to expect this by the fact that (12) is the weighted mean of the powers within the separate configurations, the weights being the probabilities $p(C)$ of the configurations. I am only concerned to show that certain methods, for which

[9] It will be noted that, when inverted, the curves of Fig. (iii) represent the power functions of tests for which the regions of rejection are those in figures (i) and (ii) respectively, the test being whether the parameter has the specified value θ, and different alternative hypotheses being represented by $(\theta + \Delta)$.

properties analogous to those of sufficiency have been claimed, do not satisfy conditions which I think they should, if these claims are to be upheld.

3. **Fiducial Distributions.** In the first section of this paper I discussed certain points of difference between the approaches to the problem of interval estimation made by R. A. Fisher on the one hand and J. Neyman and E. S. Pearson on the other. The differences are not, perhaps, of the same magnitude as those between all these writers and the protagonists of inverse probability, and the results reached are so often the same that the reader may be excused for being somewhat impatient with what appear to be rather fine distinctions. However, as was seen in the last section, the approaches do not always yield exactly the

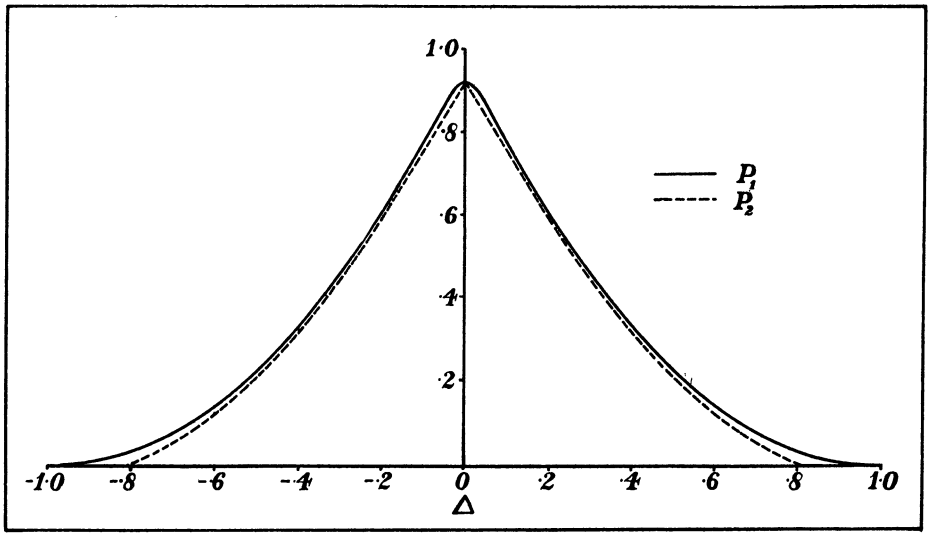

FIG. 3

same final results, and therefore I think it may be profitable to discuss them still further.

Closely connected with Fisher's desire to restrict the use of the fiducial method to situations where statistics exist which possess some property of sufficiency, is his introduction of the concept of a *fiducial distribution* for the unknown parameter. One can talk about *the* fiducial distribution for a parameter only if it is a unique distribution. Neyman, however, never makes use of fiducial distributions, and would, I think, claim that any valid results reached with the concept can equally well be reached without it. Where the results are the same there is room for two opinions on this matter. Some writers find it convenient to think in terms of fiducial distributions, and others prefer always to carry forward their reasoning as far as possible in terms of direct probability statements about the observational values, before transposing them to obtain confidence or fiducial limits for the parameters.

Greater objection can be made to the use of simultaneous fiducial distributions of several parameters. For instance, in the case of the normal distribution with parameters μ and σ, a simultaneous fiducial distribution has been defined in the following way.[10] Starting with the fact that the joint distribution of

$$\phi_1 = \frac{\sqrt{n}\,(\bar{x} - \mu)}{\sigma} \qquad \text{and } \phi_2 = \frac{(n-1)s^2}{\sigma^2}$$

is

$$df = \frac{1}{2^{\frac{1}{2}n}\Gamma(\frac{1}{2})\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{1}{2}\phi_1^2} e^{-\frac{1}{2}\phi_2} \phi_2^{\frac{1}{2}(n-3)}\, d\phi_1\, d\phi_2,$$

$\bar{x}$ and s are treated *formally* as fixed, and ϕ_1 and ϕ_2 are transformed to μ and σ, treated *formally* as variables. This gives

$$(26) \qquad df = \frac{1}{2^{\frac{1}{2}n}\Gamma(\frac{1}{2})\Gamma\left(\frac{n-1}{2}\right)} \frac{\sqrt{n}}{\sigma} e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}} \cdot \frac{2}{\sigma} e^{-\frac{(n-1)s^2}{2\sigma^2}} \left\{\frac{(n-1)s^2}{\sigma^2}\right\}^{\frac{1}{2}(n-1)} d\mu\, d\sigma$$

This distribution would be useful if it were legitimate to integrate it out to obtain a fiducial distribution for any function $g(\mu, \sigma)$ say, of μ and σ. However, as for instance Bartlett has pointed out, this is not necessarily permissible. It seems to me therefore, that distributions defined as in (26) should be dispensed with entirely, for their very form encourages the belief that they can be integrated out at will. That this belief is still held is illustrated by a recent paper by Miss D. M. Starkey[11] concerned with the difference between the means of normal populations where the standard deviations are not assumed equal. This is the original problem to which Fisher[12] applied a method equivalent to integrating out the joint fiducial distribution of the two population means. Bartlett[13] raised an objection to this method of treatment, and I have also discussed the matter further.[14] Miss Starkey proceeds from the assumption that Fisher's method is sound.

The concept of the fiducial distribution has also been used in those problems of location and scaling, which have been treated by the procedure discussed above, of considering distributions in samples with the same configuration. Indeed it is one of the attractions of this procedure that we are led to distribu-

[10] R. A. Fisher, (1935). "The fiducial argument in statistical inference." *Ann. Eugen.* VI, p. 395.

[11] Daisy M. Starkey (1937). "A test of the significance of the difference between means of samples from two normal, populations without assuming equal variances." *Ann. Math. Stat.* Vol. IX. No. 3, pp. 201-213.

[12] R. A. Fisher (1935). *loc. cit.*

[13] M. S. Bartlett (1936). "The information available in small samples." *Proc. Camb. Phil. Soc.* 32, pp. 560-566.

[14] B. L. Welch (1937). "The significance of the difference between two means when the population variances are unequal." *Biometrika*, XXIX, p. 358.

tions with, so to speak, one degree of freedom, so that the fiducial method may be safely applied. However, although probability statements based on such a fiducial method are here quite valid, I do not think that such statements can claim a *unique* validity. As I have shown in the previous section, there is no necessity to confine oneself to sampling within a configuration in order to obtain interval estimates for parameters, and we may fare better by not so confining ourselves, even if we have to dispense with the fiducial distribution.

4. Summary. Certain points which arise in the problem of estimating an interval in which a population parameter should lie have been discussed. In the second section it has been shown that in estimating location parameters it is not sufficient to consider the distribution of estimates in samples of the same configuration, meaning by sufficient that the sample is thereby utilized in the best possible way.

UNIVERSITY COLLEGE,
LONDON