SOME PROBLEMS CONNECTED WITH STATISTICAL INFERENCE

By D. R. Cox

Birkbeck College, University of London[1]

1. Introduction. This paper is based on an invited address given to a joint meeting of the Institute of Mathematical Statistics and the Biometric Society at Princeton, N. J., 20th April, 1956. It consists of some general comments, few of them new, about statistical inference.

Since the address was given publications by Fisher [11], [12], [13], have produced a spirited discussion [7], [21], [24], [31] on the general nature of statistical methods. I have not attempted to revise the paper so as to comment point by point on the specific issues raised in this controversy, although I have, of course, checked that the literature of the controversy does not lead me to change the opinions expressed in the final form of the paper. Parts of the paper are controversial; these are not put forward in any dogmatic spirit.

2. Inferences and decisions. A statistical inference will be defined for the purposes of the present paper to be a statement about statistical populations made from given observations with measured uncertainty. An inference in general is an uncertain conclusion. Two things mark out statistical inferences. First, the information on which they are based is statistical, i.e. consists of observations subject to random fluctuations. Secondly, we explicitly recognise that our conclusion is uncertain, and attempt to measure, as objectively as possible, the uncertainty involved. Fisher uses the expression 'the rigorous measurement of uncertainty'.

A statistical inference carries us from observations to conclusions about the populations sampled. A scientific inference in the broader sense is usually concerned with arguing from descriptive facts about populations to some deeper understanding of the system under investigation. Of course, the more the statistical inference helps us with this latter process, the better. For example, consider an experiment on the effect of various treatments on the macroscopic properties of a polymer. The statistical inference is concerned with what can be inferred from the experimental results about the true treatment effects. The scientific inference might concern the implications of these effects for the molecular structure of the polymer; the statistical uncertainty is only a part, sometimes small, of the uncertainty of the final inference.

Statistical inferences, in the sense meant here, involve the data, a specification of the set of possible populations sampled and a question concerning the true populations. No consideration of losses is usually involved directly in the inference, although these may affect the question asked. If the population sampled

Received October 7, 1957; revised February 10, 1958.

[1] Work done at the Department of Biostatistics, School of Public Health, University of North Carolina.

has itself been selected by a random procedure with known prior probabilities, it seems to be generally agreed that inference should be made using Bayes's theorem. Otherwise, prior information concerning the parameter of direct interest[2] will not be involved in a statistical inference. The place of prior information is discussed some more when we come to talk about decisions, but the general point is that prior information that is not statistical cannot be included without abandoning the frequency theory of probability, and information that is derived from other statistical data can be handled by methods for the combination of data.

The theory of statistical decision deals with the action to take on the basis of statistical information. Decisions are based on not only the considerations listed for inferences, but also on an assessment of the losses resulting from wrong decisions, and on prior information, as well as, of course, on a specification of the set of possible decisions. Current theories of decision do not give a direct measure of the uncertainty involved in making the decision; as explained above, a statistical inference is regarded here as having an explicitly measured uncertainty, and this is to be thought of as an essential distinction between statistical decisions and statistical inferences.

Thus, significance tests and confidence intervals, if looked at in the way explained below, are inference procedures. Discriminant analysis, considered as a method for classifying individuals into one of two groups, is a decision procedure; considered as a tool for assigning a score to an individual to say how reasonable it is that the individual comes from one group rather than the other, it is an inference procedure. Strict point estimation represents a decision; estimation by point estimate and standard error is a condensed and approximate form of interval estimation and is an inference procedure. Estimation by a posterior distribution derived from an agreed prior distribution is an inference procedure. A test of a hypothesis, considered in the literal Neyman-Pearson sense as a rule for taking one of two decisions concerning a statistical hypothesis, is a decision procedure, in which prior knowledge and losses enter implicitly. The reader may find it helpful to consider the extent to which the specification, implicitly or explicitly, of losses and prior knowledge is essential for solution of the problems just listed as ones of decision.

For example, consider the analysis of an experiment to compare two industrial processes, A and B. The statistical inference might be that, under certain assumptions about the populations, process A gives a yield higher than that of process B, the difference being statistically significant past the 1/1000 level, 90, 95 and 99 per cent confidence intervals for the amount of the true difference being such and such. The decision might be that having regard to the differences in yield of practical importance, and our prior knowledge, we will consider that the experiment has established, under the conditions examined, that process A has a higher yield than B and will take future action accordingly.

[2] *i.e.* relevant information about the parameter of interest, other than that contained in the data and in the specification of the set of possible parameter values.

An inference without a prior distribution can be considered as answering the question: 'What do these data entitle us to say about a particular aspect of the populations that interest us?' It is, however, irrational to take action, scientific or technological, without considering both all available relevant information, including for example the prior reasonableness of different explanations of a set of data, and also the consequences of doing the wrong thing. Why then, do we bother with inferences which go, as it were, only part of the way towards the final decision?

Even in problems where a clear-cut decision is the main object, it very often happens that the assessment of losses and prior information is subjective, so that it will help to get clear first the relatively objective matter of what the data say, before embarking on the more controversial issues. In particular, it may happen either that the data are little aid in deciding the point at issue, or that the data suggest one conclusion so strongly that the only people in doubt about what to do are those with prior beliefs, or opinions about losses, heavily biased in one direction. In some fields, too, it may be argued that one of the main calls for probabilistic statistical methods arises from the need to have agreed rules for assessing strength of evidence.

A full discussion of this distinction between inferences and decisions will not be attempted here. Three more points are, however, worth making briefly. First, some people have suggested that what is here called inference should be considered as 'summarization of data'. This choice of words seems not to recognise that an essential element is the uncertainty involved in passing from the observations to the underlying populations.[3] Secondly, the distinction drawn here is between the applied problem of inference and the applied problem of decision-making; it is possible that a satisfactory set of techniques for inference could be constructed from a mathematical structure very similar to that used in decision theory.

Finally, it might be argued that in making an inference we are 'deciding' to make a statement of a certain type about the populations and that therefore, provided that the word decision is not interpreted too narrowly, the study of statistical decisions embraces that of inference. The point here is that one of the main general problems of statistical inference consists in deciding what types of statement can usefully be made and exactly what they mean. In statistical decision theory, on the other hand, the possible decisions are considered as already specified.

3. The sample space. Statistical methods work by referring the observations S to a sample space Σ of observations that might have been obtained. Over Σ one or more probability measures are defined and calculations in these probability distributions give our significance limits, confidence intervals, etc. Σ is usually taken to be the set of all possible samples having the same size and structure as the observations.

[3] A referee has suggested the term 'summarization of evidence,' which seems a good one.

Fisher (see, for example, [11]) and Barnard [4] have pointed out that Σ may have no direct counterpart in indefinite repetition of the experiment. For example, if the experiment were repeated, it may be that the sample size would change. Therefore what happens when the experiment is repeated is not sufficient to determine Σ, and the correct choice of Σ may need careful consideration.

As a comment on this point, it may be helpful to see an example where the sample size is fixed, where a definite space Σ is determined by repetition of the experiment and yet where probability calculations over Σ do not seem relevant to statistical inference.

Suppose that we are interested in the mean θ of a normal population and that, by an objective randomization device, we draw either (i) with probability $\frac{1}{2}$, one observation, x, from a normal population of mean θ and variance σ_1^2 or (ii) with probability $\frac{1}{2}$, one observation x, from a normal population of mean θ and variance σ_2^2, where σ_1^2, σ_2^2 are known, $\sigma_1^2 \gg \sigma_2^2$ and where we know in any particular instance which population has been sampled.

More realistic examples can be given, for instance in terms of regression problems in which the frequency distribution of the independent variable is known. However, the present example illustrates the point at issue in the simplest terms. (A similar example has been discussed from a rather different point of view in [6], [29]).

The sample space formed by indefinite repetition of the experiment is clearly defined and consists of two real lines Σ_1, Σ_2, each having probability $\frac{1}{2}$, and conditionally on Σ_i there is a normal distribution of mean θ and variance σ_i^2.

Now suppose that we ask, accepting for the moment the conventional formulation, for a test of the null hypothesis $\theta = 0$, with size say 0.05, and with maximum power against the alternative θ', where $\theta' \simeq \sigma_1 \gg \sigma_2$.

Consider two tests. First, there is what we may call the conditional test, in which calculations of power and size are made conditionally within the particular distribution that is known to have been sampled. This leads to the critical regions $x > 1.64\ \sigma_1$ or $x > 1.64\ \sigma_2$, depending on which distribution has been sampled.

This is not, however, the most powerful procedure over the whole sample space. An application of the Neyman-Pearson lemma shows that the best test depends slightly on θ', σ_1, σ_2, but is very nearly of the following form. Take as the critical region

$$x > 1.28\sigma_1, \qquad \text{if the first population has been sampled;}$$

$$x > 5\sigma_2, \qquad \text{if the second population has been sampled.}$$

Qualitatively, we can achieve almost complete discrimination between $\theta = 0$ and $\theta = \theta'$ when our observation is from Σ_2, and therefore we can allow the error rate to rise to very nearly 10% under Σ_1. It is intuitively clear, and can easily be verified by calculation, that this increases the power, in the region of interest, as compared with the conditional test.

Now if the object of the analysis is to make statements by a rule with certain

specified long-run properties, the unconditional test just given is in order, although it may be doubted whether the specification of desired properties is in this case very sensible. If, however, our object is to say 'what we can learn from the data that we have', the unconditional test is surely no good. Suppose that we know we have an observation from Σ_1. The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution with variance σ_1^2. That is, our calculations of power, etc. should be made conditionally within the distribution known to have have been sampled, i.e. if we are using tests of the conventional type, the conditional test should be chosen.

To sum up, if we are to use statistical inferences of the conventional type, the sample space Σ must not be determined solely by considerations of power, or by what would happen if the experiment were repeated indefinitely. If difficulties of the sort just explained are to be avoided, Σ should be taken to consist, so far as is possible, of observations similar to the observed set S, in all respects which do not give a basis for discrimination between the possible values of the unknown parameter θ of interest. Thus, in the example, information as to whether it was Σ_1 or Σ_2 that we sampled tells us nothing about θ, and hence we make our inference conditionally on Σ_1 or Σ_2.

Fisher has formalized this notion in his concept of ancillary statistics [10], [23], [27]. His definitions deal with the situation without nuisance parameters and before outlining an extension that attempts to cope with nuisance parameters, it is convenient to state a slight modification of the original definitions. Let $\mathbf{m}$ be a minimal set of sufficient statistics[4] for the unknown parameter of interest, $\boldsymbol{\theta}$, and suppose that $\mathbf{m}$ can be written $(\mathbf{t}, \mathbf{a})$, where the distribution of $\mathbf{a}$ is independent of $\boldsymbol{\theta}$, and that no further components can be extracted from $\mathbf{t}$ and incorporated in $\mathbf{a}$. That is, we divide, if possible, the space of $\mathbf{m}$ into sets each similar to the sample space, and take the finest such division, assumed here to be unique subject to regularity conditions. Then $\mathbf{a}$ is called an ancillary statistic and we agree to make inferences conditionally on the observed $\mathbf{a}$.

EXAMPLES. (i) In the example of section 3, a minimal set consists of the observation, x, and an indicator variable to show which population has been sampled. The latter satisfies the conditions for being an ancillary statistic. Provided that the possible values of the mean θ include an interval, there is no set of x values with the same probability for all θ.

(ii) Under the ordinary assumptions of normal linear regression theory, plus the assumption that the independent variable has any known distribution (without unknown parameters), the values of the independent variable form an ancillary statistic.

(iii) The following example is derived from one put forward by a referee.

[4] The terms used by Fisher are that a minimal set of sufficient statistics with more components than there are parameters is called *exhaustive* and a minimal set with the same number of components as there are parameters is called *sufficient*.

Let x be a single observation with density $1 + 2\theta x$, $-\frac{1}{2} \leqq x \leqq \frac{1}{2}$, $-1 \leqq \theta \leqq 1$. Then we can write $x = [\text{sgn } x, |x|]$ and $|x|$ has the same density for all θ. Hence we argue conditionally on the observed value of $|x|$. For example in testing $\theta = 0$ against $\theta > 0$, the possible P values (see section 5) are 1 and $\frac{1}{2}$. This may seem a curious result but is, I think, reasonable if one regards a significance test as concerned with the extent to which the data are consistent with the null hypothesis.

Suppose now that there are nuisance parameters ϕ. Let $\mathbf{m}$ be a minimal set of sufficient statistics for estimating (θ, ϕ) and suppose that $\mathbf{m}$ can be partitioned into $[\mathbf{t}, \mathbf{s}, \mathbf{a}]$ in such a way that

(i) functions of $\mathbf{t}$ and θ, so-called pivotal quantities, exist with a distribution conditionally on $\mathbf{a}$ that is independent of ϕ. If any component of $\mathbf{s}$ is added to $\mathbf{t}$ or $\mathbf{a}$, this independence from ϕ no longer holds. Further, no components can be extracted from $\mathbf{t}$ and incorporated in $\mathbf{a}$;

(ii) the values of $\mathbf{a}$ and $\mathbf{s}$ give no direct information about θ in the sense to be defined below. Then we agree to make inferences about θ from the conditional distribution of (i).

We need then to define what is meant by saying that a quantity $\mathbf{y}$ gives no direct information about θ, when nuisance parameters ϕ are present. One condition that might be considered is that the density $p(\mathbf{y}; \theta, \phi)$ should be independent of θ. This seems too strong, as does also the requirement that for every different pair θ_1, θ_2 and for every $\mathbf{y}$, $p(\mathbf{y}; \theta_1, \phi) / p(\mathbf{y}; \theta_2, \phi)$ should run through all positive real values as ϕ varies. An appropriate condition seems to be that given admissible values $\mathbf{y}$, θ_1, θ_2, ϕ, there exist admissible θ, ϕ_1, ϕ_2, such that

$$\frac{p(\mathbf{y}; \theta_1, \phi)}{p(\mathbf{y}; \theta_2, \phi)} = \frac{p(\mathbf{y}; \theta, \phi_1)}{p(\mathbf{y}; \theta, \phi_2)}. \tag{1}$$

The import of the condition is that any contemplated distinction between two values of θ might just as well be regarded as a distinction between two values of ϕ.

For example, suppose that x is a single observation from a normal distribution of unknown mean ϕ and variance θ. Then x gives no direct information about θ in the sense of (1), provided that ϕ is completely unknown. Another example is normal regression theory with the independent variable having an arbitrary unknown distribution, not involving the regression parameters of interest [10]. Here $\mathbf{a}$ is the set of values of the independent variable and s is the sum squares about the regression line, assuming that the residual variance about the regression line, ϕ, is a nuisance parameter.

For a third example, let r_1, r_2 be randomly drawn from Poisson distributions of means μ_1, μ_2 and let $\mu_2 / \mu_1 = \theta$ be the parameter of interest; that is write the means as ϕ, $\phi\theta$, where ϕ is a nuisance parameter. The likelihood of r_1, r_2 can be written

$$\frac{e^{-\phi(1+\theta)}[\phi(1+\theta)]^a}{a!} \times \frac{a!}{t!(a-t)!}\left(\frac{1}{1+\theta}\right)^t\left(\frac{\theta}{1+\theta}\right)^{a-t},$$

where $t = r_1$, $a = r_1 + r_2$ and with s null. The equation (1) is satisfied, telling us that a gives us no direct information about θ. Therefore significance and confidence calculations are to be made conditionally on the observed value of a, as is the conventional procedure [25].

To apply the definitions we have to regard our observations as generated by a random process; the idea of ancillary statistics simply tells us how to cut down the sample space to those points relevant to the interpretation of the observations we have.

In the problems without nuisance parameters, it is known that methods of inference [5], that use only observed values of likelihood ratios, and not tail areas, avoid the difficulties discussed above, since the likelihood ratio is the same whether we argue conditionally or not. Lindley, using concepts from [18], has recently shown that for a broad class of problems with nuisance parameters, the conditional methods are optimum in the Neyman-Pearson sense.

Another important problem connected with the choice of the sample space, not discussed here, concerns the possibility and desirability of making inferences within finite sample spaces obtained by permuting the observations; see, for example, [16].

4. Interval estimation. Much controversy has centred on the distinction between fiducial and confidence estimation. Here follow five remarks, not about the mathematics, but about the general aims of the two methods.

(i) The fiducial approach leads to a distribution for the unknown parameter, whereas the method of confidence intervals, as usually formulated, gives only one interval at some preselected level of probability. This seems at first sight a distinct point in favour of the fiducial method. For when we write down the confidence interval $(\bar{x} - 1.96\,\sigma/\sqrt{n}, \bar{x} + 1.96\,\sigma/\sqrt{n})$ for a completely unknown normal mean, there is certainly a sense in which the unknown mean θ is likely to lie near the centre of the interval, and rather unlikely to lie near the ends and in which, in this case, even if θ does lie outside the interval, it is probably not far outside. The usual theory of confidence intervals gives no direct expression of these facts.

Yet this seems to a large extent a matter of presentation; in the common simple cases, where the upper α limit for θ is monotone in α, there seems no reason why we should not work with confidence distributions for the unknown parameter. These can either be defined directly, or can be introduced in terms of the set of all confidence intervals at different levels of probability. Statements made on the basis of this distribution, provided we are careful about their form, have a direct frequency interpretation. In applications it will often be enough to specify the confidence distribution, by for example a pair of intervals, and this corresponds to the common practice of quoting say both the 95 per cent and the 99 per cent confidence intervals.

It is not clear what can be done in those complex cases [8], [26], where say the upper 5 per cent limit for θ is larger than the upper 1 per cent limit, or indeed whether confidence interval estimation is at all satisfactory in such cases.

Within the class of distributions with monotone likelihood ratio [15], such difficulties will, however, be avoided.

If we consider that the object of interval estimation is to give a rule for making on the basis of each set of data, a statement about the unknown parameter, a certain preassigned proportion of the statements to be correct in the long run, consideration of the confidence distribution may seem unnecessary and possibly invalid. The attitude taken here is that the object is to attach, on the basis of data S, a measure of uncertainty to different possible values of θ, showing what can be inferred about θ from the data. The frequency interpretation of the confidence intervals is the way by which the measure of uncertainty is given a concrete interpretation, rather than the direct object of the inference. From this point of view it is difficult to see an objection to the consideration of many confidence statements simultaneously.

If the whole set of intervals is regarded as the fundamental concept, and if we are interested both in upper and in lower limits for θ, we may conveniently specify the set by giving say the upper and lower $2\frac{1}{2}\%$ points, etc., it being a useful convention, and no more, that the 95% interval so obtained should have equal probabilities associated with each tail. The elaborate discussion that is sometimes necessary in the conventional theory to decide which particular combination of upper and lower tail areas is best to get a 95% interval seems, from this point of view, irrelevant.

(ii) It is sometimes claimed as an advantage of fiducial estimation that it is restricted to methods that use 'all the information in the data', while confidence estimation includes any method giving the requisite frequency interpretation. This claim is lent some support by those accounts of confidence interval theory which use the words 'valid' or 'exact' for a method of calculating intervals that has, under a given mathematical set-up, an exact frequency interpretation, no matter how inadequate the intervals may be in telling us what can be learnt from the data.

However, good accounts of the theory of confidence intervals stress equally the need to cover the true value with the required probability and the requirement of having the intervals as narrow as possible in a suitable sense [21]. Very special importance, therefore, attaches to intervals based on exhaustive estimates. It is true that there are differences between the approaches in that the fiducial method takes the use of exhaustive estimates as a primary requirement, whereas in the theory of confidence intervals the use of exhaustive estimates is deduced from some other condition. This does not seem however to amount to a major difference between the methods.

(iii) The uniqueness of inferences obtained by the fiducial method has received much discussion recently, [9], [20], [28]. Uniqueness is important because, once the mathematical form of the populations is sufficiently well specified, it should be possible to give a single answer of a given type to the question 'what do the data tell us about θ?'.

The present position is that several cases are known where the fiducial method leads to non-unique answers, although it is, of course, entirely possible that a way will be found of formulating fiducial calculations to make them unique. A comparison with confidence intervals is difficult here, because in many of the multi-parameter problems, the single parameters for which confidence estimation is known to be possible at all are very limited. No cases of non-unique optimum confidence intervals seem to have been published.

(iv) If sufficient estimation, in Fisher's sense, is possible for a group of parameters, fiducial inference will usually be possible about any one of them or any combination of them, since the joint fiducial distribution of all the parameters can be found and the unwanted parameters integrated out. Exact confidence estimation is in general possible only for restricted combinations of parameters. An example is the Behrens-Fisher problem, where exact fiducial inference is possible. The situation about confidence estimation in this case is far from clear, but may be that the asymptotic expansion proposed by Welch [30], while giving a close approximation to an 'exact' system of confidence intervals, has frequency properties depending slightly on the nuisance parameters. Nothing seems to be known about possible optimum properties in the Neyman-Pearson sense. In the language of testing hypotheses, Welch's procedure is to look for a region of constant size α, independently of the nuisance parameters. It is conceivable that greater power against some alternatives is attained by having a size only bounded by α; indeed, this is made plausible by [12].

(v) The final consideration concerns the question of frequency verification. Fisher has repeatedly stated that the *immediate* object of fiducial inference is not the making of statements that will be correct with given frequency in the long run. One may readily accept this in that one really wants to measure the uncertainty corresponding to different ranges of values for θ, and it is quite conceivable that one could construct a satisfactory measure of uncertainty that has not a direct frequency interpretation. Yet one must surely insist on some pretty clear-cut practical meaning to the measure of uncertainty and this fiducial probability has never been shown to have, except in those cases where it is equivalent to confidence interval estimation. J. W. Tukey's [25] recent unpublished work on fiducial probability and its frequency verification may be mentioned here.

A different justification of fiducial distributions that is sometimes advanced is to derive them from Bayes's theorem, using a conventional form of prior distribution. To remain within the framework of the frequency theory of probability, it would then be necessary to distinguish between proper frequency distributions and hypothetical ones. The physical interpretation of the measure of uncertainty of statements about θ is that *if* θ had such and such a prior frequency distribution, then the posterior frequency distribution of θ *would be* such and such. This all amounts to a reinterpretation of Jeffreys's theory [17]. An important advantage of this approach is that it ensures independence from

the sampling rule (see [2]) and from the difficulties of section 3. On the other hand it seems a clumsy way of dealing with simple one-parameter problems, especially when the choice of prior distribution is difficult.

If the above considerations are accepted, it seems reasonable to base interval estimation on a slightly revised form of the theory of confidence intervals.

Estimation by confidence or fiducial distribution may be contrasted with the proposal [5], [13] to plot the likelihood of the unknown parameter θ in the light of the data, standardized by the maximum likelihood over θ. Advantages of the latter method are mathematical simplicity and independence from the sampling rule. Disadvantages are that it is not clear how to deal with nuisance parameters, that it is not clear that division by the maximum value of the likelihood makes values in different situations genuinely comparable, and that there is some difficulty in giving practical interpretation to the ratios so obtained. It might be argued that this last difficulty arises solely from lack of familiarity with the method.

5. Significance tests. Suppose now that we have a null hypothesis H_0 concerning the population or populations from which the data S were drawn and that we enquire 'what do the data tell us concerning the possible truth or falsity of H_0?' Adopt as a measure of consistency with the null hypothesis

$$\text{(2)} \qquad \text{prob}\left\{\begin{matrix}\text{data showing as much or more} \\ \text{evidence against } H_0 \text{ as } S\end{matrix}\,\middle|\, H_0\right\}.$$

That is, we calculate, at least approximately, the actual level of significance attained by the data under analysis and use this as a measure of conformity with the null hypothesis. The value obtained in this way is often, particularly in the biological literature, called the P-value. Significance tests are often used in practice like this, although many formal accounts of the theory of tests suggest, implicitly or explicitly, quite a different procedure. Namely, we should, after considering the consequences of wrongly accepting and rejecting the null hypothesis, and the prior knowledge about the situation, fix a significance level in advance of the data. This is then used to form a rigid dividing line between samples for which we accept the null hypothesis and those for which we reject the null hypothesis. A decision-type of this sort is clearly something quite different from the application just contemplated.

Two aspects of significance tests will be discussed briefly here. First there is the question of when significance tests are useful and secondly there is the justification of (2) as a measure of conformity.

We shall for simplicity, consider situations in which the possible populations correspond to values of a continuously varying parameter θ, the null hypothesis being say $\theta = \theta_0$. There may be nuisance parameters.

A practical distinction can be made between cases in which the null value θ_0 is considered because it divides the parameter range into qualitatively different

sections and those cases in which it is thought that there is a reasonable prospect that the null value is very nearly the true one. For example, in the comparison of two alternative industrial processes we might quite often have no particular expectation that the treatment difference is small. In such cases the significance test is concerned with whether we can, *from the data under analysis*, claim the existence of a difference in the same direction as that observed. Or, to look at the matter slightly differently, the significance level tells us at what levels the confidence intervals for the true difference include only values with the same sign as the sample difference. This idea that the significance level is concerned with the possibility that the true effect may be in the opposite direction from that observed, occurs in a different way in [17].

The answer to the significance test is rarely the only thing we should consider: whether or not significance is attained at an interesting level (say at the 10% level or better), some consideration should be given to whether differences that may exist are of practical importance, i.e. estimation should be considered as well as significance testing. A likely exception to this is in the analysis of rather limited amounts of data, where it can be taken for granted that differences of practical importance are consistent with the data. The point of the statistical analysis is in such cases to see whether the direction of any effects has been reasonably well established, i.e. whether a qualitative conclusion about the effects has been demonstrated.

The problem dealt with by a significance test, as just considered, is different from that of deciding which of two treatments is to be recommended for future use or further investigation. This cannot be tackled without consideration of the differences of practical importance, the losses consequent on wrong decisions and the prior knowledge. Depending on these and on sample size, the level of P for practical action may vary widely.

The second type of application of significance tests is to situations where there is a definite possibility that the null hypothesis is nearly true. (Exact truth of a null hypothesis is very unlikely except in a genuine uniformity trial). A full analysis of such a situation would involve consideration of what departure from the null hypothesis is considered of practical importance. However, it is often convenient to test the null hypothesis directly; if significant departure from it is obtained, consideration must then be given to whether the departure is of practical importance. Of course, in any case we will probably wish to examine the problem as one of estimation as well as of significance testing, asking for example, for the maximum true difference consistent with the data.

Consider now the choice of (2) as the quantity to measure significance. To use the definition, we need to order the points of the sample space in terms of the evidence they provide against the null hypothesis.

The most satisfactory way is the introduction, as in the usual development of the Neyman-Pearson theory, of the requirement of maximum sensitivity in the detection of certain types of departure from the null hypothesis. That is, we wish, in the simplest case, to maximise, if possible for all fixed ϵ,

$$\text{prob}_\theta(\text{attaining significance at the } \epsilon \text{ level}),$$

where θ represents a set-up which we desire to distinguish from the null hypothesis. That is we choose the procedure that makes the random variable (2) as stochastically small as possible when the alternative hypotheses are true. This leads in simple cases, to a unique specification of the significance probability (2).

In the simple case when there is a single alternative hypothesis, it seems at least of theoretical interest to distinguish between the problem of discrimination and that of significance testing. In discrimination, the two populations are on an equal footing and there are strong arguments for considering that only the observed value of the likelihood ratio is relevant. The question asked is 'which of these populations do the observations come from?' In significance testing the question is 'are the data consistent with having come from H_0 ?' The alternative hypothesis serves merely to mark out the sample points giving evidence against H_0.

The next question to consider is why we sum over a whole set of sample points rather than work in terms only of the observed point. This has been much discussed. The advantage of (2) is that it has a clear-cut physical interpretation in terms of the formal scheme of acceptance and rejection contemplated in the Neyman-Pearson theory. To obtain a measure depending only on the observed sample point, one way is to take the likelihood ratio, for the observed point, of the null hypothesis versus some conventionally chosen alternative (see [5]), and while a practical meaning can be given to this, it has less direct appeal. But consider a test of the following discrete null hypotheses:

Sample value	prob. under H_0	prob. under H_0'
0	0.80	0.75
1	0.15	0.15
2	0.05	0.05
3	0.00	0.04
4	0.00	0.01

and suppose that the alternatives are the same in both cases and are such that the probabilities (2) should be calculated by summing the probabilities of values as great or greater than that observed. Suppose further that the observation 2 is obtained; under H_0 the significance level is 0.05, while under H_0' it is 0.10. Yet it is difficult to see why we should say that our observation is more consistent with H_0 than with H_0'; this point has often been made before [4], [16]. On the other hand, if we are really interested in the confidence interval type of problem, i.e. in covering ourselves against the possibility that the 'effect' is in the direction opposite to that observed, the use of the tail area seems more reasonable. As noted in section 3 the use of likelihood ratios rather than summed probabilities avoid difficulties connected with the choice of the sample space, Σ. We are faced with a conflict between the mathematical and logical advantages of the likelihood ratio, and the desire to calculate quantities with a clear practical meaning in terms of what happens when they are calculated.

In general the role that tail areas ought to play in statistical inference is far from clear and further discussion is very desirable. The reader may refer to [1] and [19].

In this and the preceding section the problems of interval estimation and significance testing have been considered. There is not space to give a parallel discussion of the other types of statistical procedure.

6. The role of the assumptions. The most important general matter connected with inference not discussed so far, concerns the role of the assumptions made in calculating significance, etc. Only a very brief account of this matter will be given here.

Assumptions that we make, such as those concerning the form of the populations sampled, are always untrue, in the sense that, for example, enough observations from a population would surely show some systematic departure from say the normal form. There are two devices available for mitigating this difficulty, namely

(i) the idea of nuisance parameters, i.e. of inserting sufficient unknown parameters into the functional form of the population, so that a better approximation to the true population can be attained;

(ii) the idea of robustness (or stability), i.e. that we may be able to show that the answer to the significance test or estimation procedure would have been essentially unchanged had we started from a somewhat different population form. Or, to put it more directly, we may attempt to say how far the population would have to depart from the assumed form, to change the final conclusions seriously. This leaves us with a statement that has to be interpreted qualitatively in the light of prior information about distributional shape, plus the information, if any, to be gained from the sample itself. This procedure is frequently used in practical work, although rarely made explicit.

In inference for a single population mean, examples of (i) are, in order of complexity, to assume

(a) a normal population of unknown dispersion;

(b) a population given by the first two terms of an Edgeworth expansion;

(c) in the limit, either an arbitrary population, or an arbitrary continuous population (leading to a distribution-free procedure).

The last procedure has obvious attractions, but it should be noted that it is not possible to give a firm basis for choice between numerous alternative methods, without bringing in strong assumptions about the power properties required, and also that it often happens that no reasonable distribution-free method exists for the problem of interest. Thus if we are concerned with the mean of a population of unknown shape and dispersion, no distribution-free method is available [3]; when the property measured is extensive, the mean is often the uniquely appropriate parameter.

A rather artificial example of method (ii) is that if we were given a single observation from a normal population and asked to assess the significance of the difference from zero, we could plot the level attained against the population

standard deviation σ. Then we could interpret this qualitatively in the light of whatever prior information about σ was available. A less artificial example concerns the comparison of two sample variances. The ratio might be shown to be highly significant by the usual F test and a rough calculation made to show that provided that neither β_2 exceeded β_2^0, significance at least say at the 1 per cent level would still occur.

In practical situations we usually employ a mixture of (i) and (ii) depending on

(a) the extent to which our prior knowledge limits the population form in respects other than those of direct interest;
(b) the amount of information in the data about the population characteristic that may be used as a nuisance parameter;
(c) the extent to which the final conclusion is sensitive to the particular population characteristic of interest.

Thus, in (a) if we have a good idea of the population form, we are probably not much interested in the fact that a distribution-free method has certain desirable properties for distributions quite unlike that we expect to encounter. To comment on (b), we would probably not wish to studentize with respect to a minor population characteristic about which hardly any information was contained in the sample, e.g. an estimate of variance with one or two degrees of freedom. In small sample problems there is frequently little information about population shape contained in the data. Finally, there is consideration (c). If the final conclusion is very stable under changes of distribution form, it is usually convenient to take the most appropriate simple theoretical form as a basis for the analysis and to use method (ii).

Now it is very probable that in many instances investigation would show that the same answer would, for practical purposes, result from the alternative types of method we have been discussing. But suppose that in a particular instance there is disagreement, e.g. that the result of applying a t test differs materially from that of applying some distribution-free procedure. What should we do?

It can be argued that, even if we have no good reason for expecting a normal population, we should not be willing to accept the distribution-free answer unconditionally. A serious difference between the results of the two tests would indicate that the conclusion we draw about the population mean depends on the population shape in an important way, e.g. depends on the attitude we take to certain outlying observations in the sample. It seems more satisfactory for a full discussion of the data, to state this and to assemble whatever evidence is available about distributional form, rather than simply to use the distribution-free approach. Distribution-free methods are, however, often very useful in small sample situations where little is known about population form and where elaborate treatment of the results would be out of place.

An interesting discussion of the role of assumptions in decision theory is given in [14].

I am much indebted to the two referees for detailed and constructive criticism of the paper.

REFERENCES

[1] F. J. ANSCOMBE, "Contribution to the discussion of a paper by F. N. David and N. L. Johnson" *J.R. Stat. Soc.*, B, Vol. 18, (1956), pp. 24–27.

[2] F. J. ANSCOMBE, "Dependence of the fiducial argument on the sampling rule", *Biometrika*, Vol. 44 (1957), pp. 464–469.

[3] R. R. BAHADUR AND L. J. SAVAGE, "The non-existence of certain statistical procedures in non-parametric problems", *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1115–1122.

[4] G. A. BARNARD, "The meaning of a significance level", *Biometrika*, Vol. 34 (1947), pp. 179–182.

[5] G. A. BARNARD, "Statistical inference", *J.R. Stat. Soc., Suppl.*, Vol. 11 (1949), pp. 115–139.

[6] M. S. BARTLETT, "A note on the interpretation of quasi-sufficiency", *Biometrika*, Vol. 31 (1939), pp. 391–392.

[7] M. S. BARTLETT, "Comment on Sir Ronald Fisher's paper", *J.R. Stat. Soc.*, B, Vol. 18 (1956), pp. 295–296.

[8] H. CHERNOFF, "A property of some type of *A* regions", *Ann. Math. Stat.*, Vol. 22 (1951), pp. 472–474.

[9] M. A. CREASY, "Limits for the ratio of means", *J.R. Stat. Soc.*, B, Vol. 16 (1954), pp. 186–194.

[10] R. A. FISHER, "The logic of inductive inference", *J.R. Stat. Soc.*, Vol. 98 (1935), pp. 39–54.

[11] R. A. FISHER, "Statistical methods and scientific induction", *J.R. Stat. Soc.*, B, Vol. 17 (1955), 69–78.

[12] R. A. FISHER, "On a test of significance in Pearson's Biometrika Table No. 11", *J.R. Stat. Soc.*, B, Vol. 18 (1956), pp. 56–60.

[13] R. A. FISHER, *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd, 1956.

[14] W. HOEFFDING, "The role of assumptions in statistical decisions", *Proc. 3rd Berkeley Symp.*, Vol. 1 (1955), pp. 105–114.

[15] S. KARLIN AND H. RUBIN, "Theory of decision procedures for distributions with monotone likelihood ratio", *Ann. Math. Stat.*, Vol. 27 (1956), pp. 272–299.

[16] O. KEMPTHORNE, "The randomization theory of experimental inference", *J. Am. Stat. Assn.* Vol. 50 (1955), pp. 946–967.

[17] H. JEFFREYS, *The theory of probability*. Oxford, 2nd ed., 1946.

[18] E. L. LEHMANN AND H. SCHEFFE, "Completeness, similar regions and unbiased estimation, II, *Sankhya*, Vol. 15 (1955), pp. 219–236.

[19] D. V. LINDLEY, "A statistical paradox", *Biometrika*, Vol. 44 (1957), pp. 187–192.

[20] J. G. MAULDON, "Pivotal quantities for Wishart's and related distributions and a paradox in fiducial theory", *J.R. Stat. Soc.*, B, Vol. 17 (1955), pp. 79–85.

[21] J. NEYMAN, "Note on an article by Sir Ronald Fisher", *J.R. Stat. Soc.*, B, Vol. 18 (1956), pp. 288–294.

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

[23] A. R. G. OWEN, "Ancillary statistics and fiducial distributions", *Sankhya*, Vol. 9, (1948), pp. 1–18.

[24] E. S. PEARSON, "Statistical concepts in their relation to reality", *J.R. Stat. Soc.*, B, Vol. 17 (1955), pp. 204–207.

[25] J. PRZYBOROWSKI AND J. WILENSKI, "Homogeneity of results in testing samples from Poisson series", *Biometrika*, Vol. 31 (1939), pp. 313–323.

[26] C. STEIN, "A property of some tests of composite hypotheses", *Ann. Math. Stat.*, Vol. 22 (1951), pp. 475–476.
[27] J. W. TUKEY, "Fiducial inference", unpublished lectures.
[28] J. W. TUKEY, "Some examples with fiducial relevance", *Ann. Math. Stat.*, Vol. 28 (1957), pp. 687–695.
[29] B. L. WELCH, "On confidence limits and sufficiency with particular reference to parameters of location", *Ann. Math. Stat.*, Vol. 10 (1939), pp. 58–69.
[30] B. L. WELCH, "Generalisation of Student's problem", *Biometrika*, Vol. 34 (1947), pp. 28–35.
[31] B. L. WELCH, "Note on some criticisms made by Sir Ronald Fisher", *J.R. Stat. Soc.*, B, Vol. 18 (1956), pp. 297–302.