Some Problems Connected With Statistical Inference[*]

by

D. R. Cox

Department of Biostatistics, School of Public Health

Institute of Statistics
Mimeo SeriesNo. 151

[*] Invited address given at a joint meeting of the Institute of Mathematical Statistics and the Biometric Society, Princeton University, Princeton, N. J., April 20, 1956.

Some problems connected with statistical inference*

by

D. R. Cox

Department of Biostatistics, School of Public Health

1. Introduction. The aim of this paper is to make some general comments about the nature of statistical inference. Most of the points are implicit or explicit in the literature or in current statistical practice.

2. Inferences and decisions. For the present discussion a statistical inference will be defined as 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 recognize that our conclusion is uncertain and attempt to measure, as objectively as possible, the uncertainty involved.

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 essentially descriptive facts about populations to some deeper understanding of the process under investigation. The more the statistical inference helps us with this latter process the better.

Statistical inferences involve the data, assumptions about the populations sampled, a question about the populations, and very occasionally a distribution of prior probability. No consideration of losses is usually involved directly in the inference although these may affect the question asked.

Statistical decisions deal with the best action to take on the basis of statistical information. Decisions are based on not only the considerations just

*Invited address given at a joint meeting of the Institute of Mathematical Statistics and the Biometric Society, Princeton University, Princeton, N. J., April 20, 1956.

listed, but also on an assessment of the losses consequent on 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 the decision.

An inference can be considered as answering the question: "What are we really entitled to learn from these data?". A decision, however, should be based on all the information available that bears on the point at issue, including for example the prior reasonableness of different explanations of a set of data. This information that is additional to the data is called prior knowledge.

Now the general idea that we should, in any application, ask ourselves what are the possible courses of action to be taken, what the consequences of incorrect action are, and what prior knowledge is available, is unquestionably of great importance. Why, then, do we bother with inferences which go, as it were, only part of the way?

First, particularly in scientific problems, it seems of intrinsic interest to be able to say what the data tell us, quite apart from the course of action that we decide upon. Secondly, even in problems where a clear-cut decision is the sole object, it often happens that the assessment of losses and prior information is highly subjective, and therefore it may be advantageous to get clear the relatively objective matter of what the data say, before embarking on the more controversial issues.

A full discussion of this distinction between inferences and decisions will not be attempted here. Two further 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 recognize that we are essentially concerned with the uncertainty involved in passing from the observations to the underlying populations. 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 the mathematical structure used in decision theory.

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.

R. A. Fisher (see, for example, [7]) and G. A. Barnard, [2] , 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 1/2, one observation, x, from a normal population of mean θ and variance σ_1^2 or (ii) with probability 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 [4] .)

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 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, in the Neyman-Pearson sense, for the 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 $\theta', \sigma_1, \sigma_2$, but is very nearly of the following form. Take as the critical region

$x >$ 1.28 σ_1 , if the first population has been sampled

$x > 5\sigma_2$, 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. (The increase in power could be made striking by having an unequal division of probability between the two lines.)

Now if the object of the analysis is to make statements by a rule with certain desirable long-run properties, the unconditional test just given is unexceptionable. If, however, our object is to say 'what we can learn from the data that we have', the unconditional test is surely unacceptable. 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 been sampled, i.e. we should use the conditional test.

To sum up, in statistical inference 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 was 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 [6][11]. As orignially put forward, this seems insufficiently general to deal with such situations as the 2 x 2 contingency table, and the following generalization is put forward. Suppose that we wish to make an inference about parameters $\underset{\sim}{\theta}$, with parameters $\underset{\sim}{\phi}$ regarded as nuisance parameters and that exhaustive estimation of ($\underset{\sim}{\theta}$, $\underset{\sim}{\phi}$) can be based on functions $\underset{\sim}{t}$, $\underset{\sim}{a}$, $\underset{\sim}{s}$ of the observations, such that

(i) the distribution of $\underset{\sim}{t}$ given $\underset{\sim}{a}$ depends only on $\underset{\sim}{\theta}$;

(ii) the values of $\underset{\sim}{a}, \underset{\sim}{s}$ give no information about $\underset{\sim}{\theta}$ in that their joint distribution function $p(\underset{\sim}{a}, \underset{\sim}{s}; \underset{\sim}{\theta} \; \underset{\sim}{\phi})$ is such that for any $\underset{\sim}{a}, \underset{\sim}{s}, \underset{\sim}{\theta}, \underset{\sim}{\phi}, \underset{\sim}{\theta},$ there exists $\underset{\sim}{\phi}_1$ such that

$$p\,(\underset{\sim}{a}, \underset{\sim}{s}\,; \underset{\sim}{\theta}, \underset{\sim}{\phi}_1) = p\,(\underset{\sim}{a}, \underset{\sim}{s}\,; \underset{\sim}{\theta}, \underset{\sim}{\phi}); \qquad *$$

then inference about $\underset{\sim}{\theta}$ should be based on the distribution of $\underset{\sim}{t}$ given $\underset{\sim}{a}$, i.e. the sample space Σ should consider $\underset{\sim}{a}$ as fixed and equal to its observed value.

To apply this definition we have to regard our observations as generated by a random process; the definition simply tells us how to cut down the sample space to those point relevant to the interpretation of the observations we have. The equation $*$ is the formal expression of the condition that $\underset{\sim}{a}$ (and $\underset{\sim}{s}$) give no basis for discrimination between different values of $\underset{\sim}{\theta}$.

For 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

$$\frac{e^{-\phi(1+\theta)}\,[\phi(1+\theta)]^{a}}{a!} \times \frac{a!}{t!\,(a-t)!}\left(\frac{1}{1+\theta}\right)^{t}\left(\frac{\theta_1}{1+\theta}\right)^{a-t},$$

where $t = r_1$, $a = r_1 + r_2$. The equation $*$ is easily shown to be satisfied, telling us that a gives us no information about θ . Therefore significance and confidence calculations are to be made conditionally on the observed value of a, as is the conventional procedure [12] . Note that if the populations for study were selected by a random procedure independent of θ , the likelihood would be changed only by a factor independent of θ and the final choice of sample space would be unchanged.

A method of inference that used only the values of likelihood ratio would avoid these difficulties [3] .

Another important problem connected with the choice of the sample space concerns the possibility and desirability of making inferences within finite sample spaces obtained by permuting the observations . This matter will not be discussed here.

4. Interval estimation. Much controversy has centered 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 favor of the fiducial method. For when we write down a confidence interval such as $(\bar{x} - 1.96\,\sigma/\sqrt{n},\ \bar{x} + 1.96\,\sigma/\sqrt{n})$, there is certainly a sense in which the unknown mean θ is likely to lie near the center 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; 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.

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 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. It is then difficult to see an objection to the consideration of many confidence statements simultaneously.

As an example, consider the estimation of the mean θ of a normal population of unit variance, from a single observation x, when it is given that $\theta \geq 0$. The natural confidence distribution is

at $\theta = 0$, a point probability, $\frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{-x} e^{-\frac{1}{2}t^2} dt$,

for $\theta > 0$, a continuous distribution $\frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}(x-\theta)^2}$

The frequency interpretation of this is that if we were to select always the set of θ values covered by some fixed part of the confidence distribution, e.g. the lower 95 per cent, then for any $\theta > 0$, the true value is covered in 95 per cent of trials in the long run. For $\theta = 0$, the corresponding frequency exceeds 95 per cent, or can be made equal to 95 per cent by a natural process of random selection. Such a random process to achieve exactly 95 per cent coverage is a process to explain the concrete meaning of the confidence distribution; it does not seem relevant to the actual use of the distribution in applications.

If the restriction is that $\theta > 0$ there is a difficulty in that $\theta = 0$ is an inadmissable parameter value and it is not sensible to attach confidence probability to a parameter value that cannot occur. This seems, however, a rather artifical matter and it seems reasonable to deal with it by the convention that $\theta = 0$ is to stand for some very small positive value of θ.

(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. Very special importance, therefore, attaches to intervals based on exhaustive estimates. It is true that there are differencesbetween 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. There does not seem a major difference between the methods here.

(iii) The uniqueness of inferences obtained by the fiducial method has received much discussion recently, [5], [10], [13] . Uniqueness is important because, once the mathematical form of the populations is sufficiently well specified, it should be possible to give a single answer to the question 'what do the data tell us about Θ'. Yet too much must not be made of this issue because of the doubts that always arise as to the appropriate theoretical specification.

The present position is that several cases are known where the fiducial method leads to non-unique answers and, so far as I know, none where the confidence intervals, based on exhaustive estimates, are not unique. It is, of course, entirely possible that a way will be found of formulating fiducial calculations to make them unique. But it is clear that at present there is no ground for preferring the fiducial approach on these considerations.

(iv) If exhaustive estimation 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 not too clear but is probably that the procedure preferred by Welch, while giving a close approximation to an 'exact' system of confidence intervals, has frequency properties depending slightly on the nuisance parameters.

This point is, the possibility that no exact confidence interval solution to the problem exists was recognized by Welch: see also [14]. Somewhat academic is that these considerations are all with respect to those idealized conditions in which sufficient statistics exist, however, within this framework the fiducial technique is certainly the more powerful.

(v) The final consideration concerns the question of frequency verification. R. A. Fisher has repeatedly stated that the object of fiducial inference is not to make statements that will be correct with given frequency in the long run. One may agree with 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 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. J. T. Tukey's recent work on fiducial probability and its frequency verification should be referred to here.

It seems, therefore, that with some shifts of emphasis, the theory of confidence intervals is adequate to deal with the problem of the interval estimation.

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

$$\mathrm{prob}_{\Sigma}\left\{\begin{matrix}\text{data showing as much}\\ \text{or more evidence against } H_0 \text{ as } S\end{matrix}\;\middle|\; H_0\right\} . \quad (**)$$

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. Significance tests are often used in practice in this way, although many formal accounts of the theory of tests suggest, implicitlyor 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 test of this sort is clearly something quite different from the application contemplated here.

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 (**) as a measure of conformity. The discussion is restricted to the testing of simple hypotheses about unknown parameters, with or without nuisance parameters. For example, if θ is the mean of a normal population, we consider tests of $\theta = 0$, but not of $\theta < 0$.

Perhaps the most frequent type of application of significance tests, at any rate in technological work, is in situations where the null hypothesis is almost certainly false and where, moreover, we have no particular reason to think that it is even approximately true. For example, in the comparison of two alternative industrial processes we would usually be certain that an experiment of sufficient sensitivity would show there to be some real difference between the processes in whatever property is of interest. The significance test is concerned with whether we can, from the data under analysis, claim the existence of a difference. 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 [9].

Hardly ever is the answer to the significance test the only thing we should consider: whether or not significance is attained at an interesting level (say at least at the 10% level), some consideration should

* F. J. Anscombe [1] has recently given a very interesting discussion of this.

be given to whether differences that may exist are of pratical importance, i.e. estimation should be considered as well as significance testing. A possible exception to this is in the analysis of very limited amounts of data. Here it can often be taken for granted that differences of pratical importance are consistant with the data, the point of the statistical analysis being to see whether the direction of any effects has been reasonably well established.

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. This cannot be tackled without careful consideration of the differences of practical importance, the losses consequent on wrong decisions and the prior knowledge.

Another type of application of significance tests is to situations where there is a definite possibility that the null hypothesis is very nearly true. (Exact truth of a null hypothesis is 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, we probably in any case will wish to examine the problem as one of estimation as well as of significance testing.

Consider now the choice of (**) 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. There are two ways of doing this. The first, and most satisfactory, 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 simpliest case, to maximize, if possible for all fixed θ, ε,

$$\text{prob}_{\Theta} \quad (\text{attaining significance at the } \varepsilon \text{ level}),$$

where Θ represents a set-up which we derive to distinguish from the null hypothesis. This leads, in simple cases, to a unique specification of the significance probability (**).

A second method of ordering the sample points to determine (**) ; which leads to a unique answer for discrete distributions, involves the null hypothesis itself and uses no appeal to the notion of alternative hypotheses. We say that sample point S_1 shows as much or more evidence against H_o than the sample point S_2 if only

$$\text{prob}_{\Sigma} \; (S_1 \mid H_o) \leq \text{prob}_{\Sigma} \; (S_2 \mid H_o),$$

and hence calculate (**) by summing over all points with probability, under the null hypotheses, less than or equal to that of the observed point. We may call this a test of pure consistency with the null hypothesis as opposed to the previous type, which we may call a test of specific discrimination. It is clear that if we are in a position to specify what type of alternative we wish to detect, it will be much better to use the first type of test. In some standard cases, the two methods give identical answers.

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 discussed. The advantage of (**) 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, it seems necessary to take the likelihood ratio, for the observed point, of the null hypothesis versus some conventionally chosen alternative (see [3]), and the practical meaning that can be given to this is much less clear. But consider a test of the following discrete null hypothesis H_o, H_o'

Sample value	prob. under H_o	prob. under H_o'
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 (**) should be calculated by summing probabilities over the upper tails of the two distributions. Suppose further that the observation 2 is obtained; order H_o the significance level is 0.05, while under H_o' it is 0.10. Yet it is difficult to see why we should say that our observation is more connected with H_o' than with H_o; this point has often been made before, [2] , [9] . 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 §3 the use of likelihood ratios rather than summed probabilities avoids 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 the methods are used.

Further discussion of this is necessary.

6. Other questions about populations. The preceding sections have dealt briefly with inference procedures for interval estimation and for significance tests. There are numerous other questions that may be asked about the populations sampled and it would be of value to have inference procedures for answering them. For example there are the problems of selection, e.g. that of choosing from a set of treatments, a small group having desirable properties. Decision solutions of this and other similar problems are known; methods that measure the uncertainty connected with these situations do not seem available.

Again the problem of discrimination, i.e. of assigning an individual to one of two (or more) groups, is usually answered as a decision problem; that is we specify a rule for classifying a new individual into its appropriate group (we may include a 'doubtful' group as one possible answer). An inference solution would measure the strength of evidence in favor of the individual being in one or other group. The natural way to do this seems to be to quote the (log) likelihood ratio for group I versus group II.

7. 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; I do not feel competent to give the question the searching discussion that it deserves.

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 overcoming this difficulty.

(i) the idea of nuisance parameters, i.e. of inserting sufficient unknown parameters into the functional form of the population, so that a good 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 reference 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, an arbitrary population (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 difference between the means of two populations of different and unknown shapes and dispersions, no distribution-free method is known that is not palpably artificial. For these reasons, and others - for example the exemption of a dependence - distribution-free methods are not a full solution of the difficulty.

An 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 the 1 per cent level would still occur.

The choice between methods (i) and (ii) depends on

(a) the extend to which our prior knowledge limits the population from;

(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 population characteristic about which hardly any information was contained in the sample, e.g. as 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 formas 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 were to differ materially from that of applying some distribution-free procedure. What would we do?

It seems to me that, even if we have no good reason for expecting a normal population, we would not be willing to accept the distribution-free answer unconditionally. A serious difference between the results of the two tests would usually 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 to simply 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 discussion of the results would be out of place.

Clearly much more discussion of these problems is needed.

REFERENCES

[1] Anscombe, F. J., "Contribution to the Discussion of a Paper by F. N. David and N. L. Johnson," J. R. Statist. Soc., B, to appear.

[2] Barnard, G. A., "The Meaning of a Significance Level," Biometrika 34 (1947), 179-182.

[3] Barnard, G. A., "Statistical Inference," J. R. Statist. Soc., Suppl. 11 (1949), 115-139.

[4] Bartlett, M. S., "A Note on the Interpretation of Quasi-Sufficiency," Biometrika 31 (1939), 391-392.

[5] Creasy, M. A., "Limits for the Ratio of Means," J. R. Statist. Soc., B, 16 (1954), 186-194.

[6] Fisher, R. A., "The Logic of Inductive Inference," J. R. Statist. Soc. 98 (1935), 39-54.

[7] Fisher, R. A., "Statistical Methods and Scientific Induction," J.R. Statist. Soc., B, 17 (1955), 69-78.

[8] Kempthorne, O., "The Randomization Theory of Experimental Inference," J. Am. St. Assoc. 50 (1955), 946-967.

[9] Jeffreys, H., The Theory of Probability. Oxford, 2nd ed., 1946.

[10] Mauldon, J. G., "Pivotal Quantities for Wishart's and Related Distributions and a Paradox in Fiducial Theory," J. R. Statist. Soc., B, 17 (1955), 79-85.

[11] Owen, A. R. G., "Ancillary Statistics and Fiducial Distributions," Sanktya 9 (1948), 1-18.

[12] Przyborowski, J. and Wilenski, H., "Homogeneity of Results in Testing Samples from Poisson Series," Biometrika 31 (1939), 313-323.

[13] Tukey, J. W., "Fiducial Inference," unpublished lectures.

[14] Wilks, S. S., "On the Problem of Two Samples from Normal Populations With Unequal Variances," Ann. Math. Statist. 11 (1940), 475 (abstract).