Significance Tests - a discussion BOB GILCHRIST November 1976

This is a personal account of a meeting of the Radstats teaching subgroup and so, although it attempts to cover the remarks of other contributors, it is naturally enough a biased account. Dave Jarrett is adding his ideas in an accompanying article.

As our discussions developed it became clear that there was a general feeling that significance tests had achieved an overemphasis in journals. (A recent example of this phenomenon is given by Chatfield, 1976). It was felt that the over-importance of significance tests is particularly prevalent in the fields of medicine and sociology, where editors have come to regard significance tests as the measure of the respectability of an article. (It has even been suggested by Bross, 1971, that in fields where there is a serious chance of publishing effects which are not clearly established by data, the attainment of a significance level of, say, 0.05 is a reasonable requirement for the publication of results). Indeed, it was noted by the discussants that some journals were often not interested in non-significant evidence; a position complemented by the failure of many such journals to accept purely descriptive statistics. The historical development of this position was later discussed, particularly with respect to sociology, and it was noted that researchers might feel tempted to carry out significance tests as a precaution against others doing so - a sort of self-preservation which had developed into editorial convention. At the same time it was suggested that researchers had found in significance tests an automatic form of induction, and indeed that statistical tests had offered the respectability of scientific methodology to the emerging **social sciences.**

In considering the methodology of significance tests, it was noted that abuse was rife. Applied workers sometimes looked to statistics to '**prove**' the validity of their subjective judgements, whereas significance tests were often no match for the experience and knowledge of the researcher. Indeed, in a real-life situation, informal techniques could often be so clearcut that calculations of any probability level should not be necessary, although applied workers seem reluctant to accept this position. Some researchers were also not fully aware of the limitations of the statistical techniques which they used. Perhaps of greater importance was the wider misunderstanding of tests; for example, their use to 'prove' a substantive difference in a census. Or the uncritical use of tests without regard to underlying models, illustrated by the lack of data transformations. At the same time, it can be argued that applied workers might not regard significance tests in the rigorous framework of a pure statistician, but would perhaps merely see them as a descriptive measure, perhaps to test the 'randomness' of the data. (an idea due to Fisher?). As so used, significance tests seemed less objectionable, even when used in doubtful circumstances. However, most applications of significance tests would probably be within a framework of testing a null-hypothesis against some alternative. (As an aside, it may be worth reflecting at this point that the original development of the χ^2 significance test by Karl Pearson, 1900, was for the case of no alternative hypothesis - he used an absolute test with sample points ordered in decreasing probability under the null-hypothesis: see also Martin-Lof, 1975 . However, difficulties arise if we try to apply such tests to a continuous situation. In any case, it would seem that the occasions when an absolute test is sensible are very rare in applications: see Cox, 1976).

The discussion of the role of the hypothesis brought forward the suggestion that the hypothesis formulation could be particularly unsuitable for applied research, because of the lack of a direct indication of the importance of any departure from a null-hypothesis. Thus, in sociology, the null-hypothesis could be one in which the researcher did not believe; e.g. a zero correlation. Collection of sufficient data would then tend to reject

the null-hypothesis, through observing a non-zero (although substantively insignificant) correlation. A second possibility is that substantive differences might be so delicate that the null-hypothesis could be perpetually accepted as being true - although the data may also be consistent with departures from H_0 which are of great practical significance. The latter possibility obviously leads to an acceptance of the status-quo through 'statistical methods'. It is also well known that large significance levels do not necessarily coincide with large values of similar measures of association (Duggan and Dean, 1968), although such measures are not without many methodological problems of their own. In further discussions of the role of such tests it may prove useful to apply the classification given by Cox, 1976. (This paper was unfortunately not available before our discussions). He considers that there are two main categories of null-hypothesis; the plausible and the dividing. The plausible hypothesis can come in two forms - primary or simplifying. The primary plausible null-hypothesis is of intrinsic interest; he quotes the example of whether data is consistent with the hypothesis of a random walk. A simplifying plausible hypothesis can be of simple primary structure - e.g. does a variance - covariance matrix offer a convenient, complete summary of some multivariate data; or the hypothesis can be of simple secondary structure - e.g. is some two sample multivariate data sufficiently normally distributed to allow the use of Hotelling's T^2 to test equality of mean vectors. Rejection of a simple primary structure would lead to a whole new model whereas rejection of simple secondary structure would leave the hypothesis of equal means unchanged, but would perhaps necessitate an alternative statistic. Dividing hypothesis divide the range of possibilites into qualitatively different types; thus $\mu_1 = \mu_2$ divides the situations with $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$. Cox considers that this sort of hypothesis is of legitimate interest, even if not plausible, saying that if the data are reasonably consistent with the null-hypothesis then the data by themselves give no clear indication of the sign of $\mu_1 - \mu_2$. For example, the hypothesis that a point process is a Poisson process is used for dividing 'over-dispersion' from 'under-dispersion' even if the null-hypothesis is not plausible. In the context of our earlier remarks, acceptance of this null-hypothesis would merely lead to the conclusion that the data is inadequate to notice the direction of the departure, in the sense tested.

During the discussion of possible legitimate uses of significance tests, it was concluded that there were legitimate cases, although some might argue that many (all?) such cases should really be couched in a decision theory framework; e.g. the testing of, say, the hardness of two physical compounds. The sequential fitting of linear models was also put forward as an acceptable example of testing, although the choice of 'appropriate' significance levels could present problems. Nevertheless, there seemed few examples where the main summary of an analysis should be a significance test. In his recent paper, Cox suggests that tests can form the main summary of data when (1) there is a plausible null-hypothesis of intrinsic interest and (2) there is so little data that it can be assumed both that (a) evidence of inconsistency corresponds to a departure of scientific importance and (b) that even if the data agree very well with the null-hypothesis they are also consistent with departures of scientific importance. As an alternative to (2) he suggests a situation where instead, (2)' the data are so extensive that it is reasonable to assume that consistency with the null-hypothesis implies an absence of any effect of practical importance and (3) a reasonably high observed significance level is obtained, (i.e. accepting H_0). He considers that significance tests are also acceptable as a central conclusion where (1) there is a dividing null-hypothesis of the absence of structure and (2) there is such a limited amount of data that it can be assumed that data consistent with the null-hypothesis are consistent also with departures of scientific importance and (3) a reasonably large observed significance level is obtained. (again accepting H_0). Cox further suggests

that in complex situations, with plausible hypotheses of simple primary structure, simplifying assumptions are essential for incisive interpretation. Thus he considers that tests can be used, for instance, to examine data for linearity, absence of interaction, or parallel regression lines where these are not the primary objects of the analysis, so that the object of these tests is to find a simple formulation for final analysis. At the same time he also states that significance tests of simple secondary structure are generally best avoided, if possible. To the author it appeared that our discussions had much in common with this sort of approach, although a lot of additional questions were also raised.

Coming back more specifically to the discussions it seems worthwhile to mention that confidence intervals were not covered explicitly. However, although the use of interval estimation is evidently preferable to point estimation, the basis of confidence intervals seems intrinsically linked to the use of significance tests with fixed levels. The spirit of any acceptance of significance tests by the author is not to regard any fixed conventional probability level α , nor indeed to deal quite differently with cases for which the observed $p \leq \alpha$ or for which $p > \alpha$. Thus, even if significance levels are used, then 0.1, 0.05, 0.01, etc., are seen merely as convenient points of reference. This position seems almost to rule out the use of confidence intervals as anything but a crude rule of thumb.

The theme of repeated experiments also arose during our discussions. Even if a replicated experiment is costly and takes a long time, it is nevertheless possible to form independent exploratory and confirmatory experiments by splitting the data at random into two parts, (apparently to the disapproval of Fisher), using the second part to assess the significance of results suggested after exploratory analysis of the first part. Such a procedure might be advocated where the researcher did indeed wish to give importance to a test of significance, although there is a problem that different splits can of course give different conclusions. This matter also leads naturally to the consideration of other problems caused by modifying analysis in the light of data. One school of thought advocates always carrying both an exploratory and confirmatory analysis. In his recent paper, Cox also discusses this matter and the related need for making 'an allowance for selection' when defining a significance level whose hypothetical physical interpretation is directly related to the analysis which was carried out. This seems an interesting topic for our future considerations.

Bearing in mind the many issues raised and the complexity of the matter, it might be reasonably concluded from our discussions that, although it did not seem that significance tests are totally too dangerous to use, they are perhaps too dangerous to teach, at least without a warning of their pitfalls. But, unfortunately, a series of earlier discussions led to many of us having the unhappy feeling that, with the limited time available in many non-statistics degree courses, it is just not possible to find time to teach both methodology and an understanding of the underlying philosophy. And it seems almost impossible to teach the understanding without the methodology. So we are left in a quandary.

In conclusion, we should remark that our discussants did not put forward any strong Bayesian views so our coverage of this was scanty. Of course, adherents of the coherence school (de Finetti, 1975) could not accept significance tests; but it would be interesting to know their reactions to the provocative view which was put forward in our discussions, namely that Bayesian analysis often appeared to offer more than a significance test but that, in reality, Bayesian analysis offered another form of automatic inference and, consequentially, could have many of the problems inherent in significance testing. (see Bakan, 1967)
Perhaps we can discuss this further in a later edition of RSJ?

REFERENCES

BAKAN, D. (1967). On Method. San Francisco: Josey-Bass.

BROSS, I.D.J. (1971). Critical levels, statistical language and scientific inference (with discussion). In: Foundations of statistical inference, pp 500-519, eds. Godambe, V.P. and Sprott, D.A. Toronto: Holt, Rinehart and Winston.

CHATFIELD, C. (1976). A statistical true story. BIAS, 3, 2, 1-18.

COX, D.R. (1967). The role of significance tests. Paper presented to the European Meeting of Statisticians. Grenoble.

de FINETTI, B. (1975). Theory of Probability, A Critical Introductory Treatment. Vol 1 and 2. (translated from the Italian by Antonio Machi and Adrian F.M. Smith), London and New York, Wiley.

DUGGAN, T.J. and DEAN, C.W. (1968). Common misinterpretations of significance levels in sociological journals. The American Sociologist, 3, 45-46.

MARTIN-LOF, P. (1976). Reply to Sverdurp's polemical article Tests without power. Scand. J. Statist. 2, 161-165.

PEARSON, K. (1900). On the criteria that a given system of deviations from the probable in a system of variables is such that it can reasonably be supposed to have arisen from random sampling. Phil. Mag., Series 5, 50, 157-175.

Note: The above papers by Duggan and Dean, and by Bakan appear in HENKEL, R.E. (1970). The Significance Test Controversy, London: Butterworths.

SIGNIFICANCE TESTS

David Jarrett, Middlesex Polytechnic

I want to consider criticisms of the use of significance tests in the social sciences and, in particular, question the value of the conventional introductory statistics course for social science students. The kind of course I have in mind has as its main aim an introduction to statistical inference; after a fairly brief treatment of descriptive statistics nearly all the methods included are significance tests, comparatively little attention being paid to estimation. Statistical inference is justified in terms of scientific induction and opposed to "merely descriptive" statistics, which is considered trivial and unimportant, with only those topics essential for understanding the inference part of the course being included. This emphasis on significance testing reflects that of much of the social science literature - the practice of some journals of publishing only those papers which report results attaining a certain level of significance is well known (Sterling, 1959) - but many authors have been critical of the use of significance tests. In sociology and psychology this has led to a controversy which has been collected and summarised by Henkel and Morrison (1970). Doubts about the value of significance testing have also been raised in geography (Gould, 1970); economists are usually more interested in estimation and do not use significance tests to the same extent as other social scientists.

Statisticians themselves, of course, are not in agreement about significance testing. I do not want to get involved in controversies about the foundations of statistical inference, but I think it is fair to say that in most introductory courses (and in most applications in the behavioural sciences) the approach is somewhere between that of Fisher (testing a null hypothesis without a specific alternative: if the probability of getting the observed or a more extreme result is less than a certain amount, then "either an exceptionally rare chance has occurred, or the theory of random distribution (the null hypothesis in his example) is not true" - Fisher, 1956) and that of Neyman, Pearson and Wald (a decision is made between a null hypothesis and a specified alternative hypothesis, the test used being justified by its optimal long run properties). Both approaches are criticised by statisticians of the Bayesian and likelihood schools. Some of the papers in the Henkel-Morrison collection echo the Fisher-Neyman controversy of the thirties, while others (e.g. Bakan, 1967) recommend increased use of Bayesian methods.

Henkel and Morrison draw a distinction between statistical and philosophy of science issues in the use of significance tests. Some of the statistical issues are technical or concern researchers' misunderstanding of the tests, but I think that others are difficult to separate from the philosophy of science issues; these are concerned with the problem of whether the use of significance tests is appropriate in the creation of scientific knowledge. Therefore, I will not attempt to discuss these issues in a general way but will concentrate on the role of significance testing in assessing theories.

Suppose we have a theory that two variables X and Y are related (for instance there is a non-zero correlation ρ between X and Y), and wish to test this theory. A good textbook (such as Blalock, 1972) instructs us to proceed as follows: we set up the null hypothesis that X and Y are not related ($H_o: \rho = 0$), collect data (a random sample from the population) and carry out the test; if the null hypothesis is rejected at a sufficiently small significance level we can conclude that in fact X and Y are related. The criticism of this procedure is simple: there are usually a priori reasons for believing a (point) null hypothesis to be false - only rarely would we seriously consider the hypothesis that the correlation between X and Y was exactly zero - so, since most standard tests have asymptotic power one, the null hypothesis will always be rejected (at any significance level) if we take a large enough sample; in other words, our theory will always be corroborated. (One-sided tests - $H_o: \rho \leq 0$ against $H_1: \rho > 0$ - are a little more complicated; perhaps we could invoke the principle of indifference and conclude that the probability of corroborating our theory is 50% - see Meehl, 1967.)

The near certainty of rejecting the null hypothesis is a well known danger of such tests and it will be argued that the significance level should be chosen carefully, taking account of the power function of the test. There is little evidence, however, that many social scientists do this - standard levels of 5% and 1% are often preferred, and details of the power functions of standard tests are not taught in elementary courses, presumably because of technical difficulties. We can also ask for more precisely formulated hypotheses or for greater emphasis on estimation rather than testing - the existence of a deviation from the null hypothesis is rarely in doubt; what is important is its size. However, the testing of a priori false null hypotheses against vague alternatives is often justified on the grounds that the theory being tested is imprecisely formulated (Henkel and Morrison state that most hypotheses in behavioural science are "atheoretical in several ways") so perhaps the greatest need is for more developed theories.

Not all tests fit into the above framework - sometimes the null hypothesis itself is of interest and is being tested with the hope of acceptance rather than rejection. I think that at least two cases can be distinguished here: the hypothesis may be precise quantitative prediction (e.g. of the velocity of light in beer), or it may offer a useful approximation, perhaps stating that the relationship between two variables is linear, or that a series of events occur in a Poisson process. More credence might be attached to the hypothesis in the first case than in the second, but in neither instance would we seriously consider it to be exactly true (Rubin, 1971, notes as possible exceptions the constancy of the velocity of light in a vacuum and the non-existence of extra-sensory perception) so again the null hypothesis is certain to be rejected if enough data is available; but now our theory is always refuted. Rubin concludes that the use of tests at a fixed level of signficance is not appropriate. The second case is better regarded as an estimation, rather than a testing, problem - if the null hypothesis of linearity is accepted then we can use a simpler estimation procedure than if we decide that non-liniear regression is necessary - and introduces the problem of preliminary testing, currently a fashionable topic in the econometrics literature (e.g. Judge, Bock and Yancey, 1974). Clearly the choice of signficance level is of crucial importance here, and recent work of Leonard (1976) has shown that Bayesian methods can lead to the simpler estimation procedure (acceptance of the null hypothesis in classical terms) even though a significance test would reject the null hypothesis at any sensible significance level.

A further objection to the use of tests of significance concerns the problem of defining the population to which the inference applies. The theory of statistical inference is founded on probability theory, and classical procedures require a probability model for the data; in special cases the probability model is introduced artificially by selecting a sample at random from a real, finite population. Scientific theories are not usually concerned with finite populations, yet in the social sciences (except econometrics and other areas where more sophisticated stochastic models are used) the probability model is rarely stated explicitly; the problem becomes particularly acute when inferences are made from official statistics or from census data. Many elementary courses justify the procedures of statistical inference in the context of sampling from a finite population, with a hand-waving extension to a hypothetical, infinite population. Bakan (1967) makes the additional point that theories in psychology usually concern individuals, whereas significance tests only enable us to make conclusions about aggregates; a further induction from the aggregate to the general is necessary in order to obtain meaningful scientific propositions. The failure to appreciate this is part of a general confusion between statistical and scientific inference; major developments in the physical sciences took place without the use of significance tests, but to the behavioural scientist the tests appear to offer an automatic form of scientific induction.

Bakan concludes:

"What we have indicated in this paper in connection with the test of significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research which may be related to the presumption of the non-existence of mind in the subjects of psychological research."

I believe that much of what we teach is of limited value if we want to encourage meaningful research in the social sciences. Moreover, Ehrenberg (1976) claims that what is taught is irrelevant for the everyday application of statistics. Perhaps we succeed only in giving our students the impression that statistics is a substitute for thinking. I do not know a universal solution to the problem but can suggest some improvements. Within the framework of the conventional course we can stress the limitations of the theory and attempt to make sure that the probability basis for inference is understood (though the latter may be difficult with students who do no mathematics). Certainly the treatment of descriptive statistics (understood as a body of techniques for exploring and interpreting data) can be extended. Some multivariate data-analytic techniques such as cluster analysis can be taught in an elementary course, though it might be objected that the unthinking use of such methods is as dangerous as the unthinking use of significance tests. Social scientists have been persuaded that statistical inference is important - it is taught to virtually all social science students, although they are not necessarily taught mathematics or the philosophy of science; I think that the philosophy of science in particular is at least as important as statistics.

REFERENCES

BAKAN, D. (1967), The test of significance in psychological research, from On Method, San Francisco: Jossey-Bass, 1-29. (Reprinted in Henkel and Morrison, 1970)

BLALOCK, H.M. (1972), Social Statistics, McGraw-Hill

EHRENBERG, A.S.C. (1976), We must teach what is practised, The Statistician, Vol. 25, No. 2

FISHER, R.A. (1956), Statistical Methods and Scientific Inference, Edinburgh: Oliver and Boyd

GOULD, P. (1970), Is statistix inferens the geographical name for a wild goose? Economic Geography, Vol. 46 (2), 439-50

HENKEL, R.E. and MORRISON, D.E. (1970), The Significance Test Controversy, London: Butterworths

JUDGE, G.G., BOCK, M.E. and YANCEY, T.A. (1974), Post data model evaluation, Review of Economics and Statistics, Vol. LVI, 245-53

LEONARD, T. (1976), The Bayesian analysis of categorical data, Paper presented to the University of London Joint Statistics Seminar, 29 October, 1976

MEEHL, P.E. (1967), Theory testing in psychology and physics: a methodological paradox, Philosophy of Science, Vol. 34, 103-15 (Reprinted in Henkel and Morrison, 1970)

RUBIN, H. (1971), Occam's razor needs new blades, in Godambe, V.P. and Sprott, D.A. (eds.), Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston

STERLING, T.D. (1959), Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa, Journal of the American Statistical Association, Vol. 54, 30-34 (Reprinted in Henkel and Morrison, 1970)