A STATISTICAL PARADOX

By D. V. LINDLEY

Statistical Laboratory, University of Cambridge

An example is produced to show that, if H is a simple hypothesis and x the result of an experiment, the following two phenomena can occur simultaneously:

(i) a significance test for H reveals that x is significant at, say, the 5 % level;

(ii) the posterior probability of H, given x, is, for quite small prior probabilities of H, as high as 95 %.

Clearly the common-sense interpretations of (i) and (ii) are in direct conflict. The phenomenon is fairly general with significance tests and casts doubts on the meaning of a significance level in some circumstances.

We begin by giving the mathematical derivation of the example and later comment on it and the assumptions involved. Let $(x_1, x_2, \ldots, x_n)$ be a random sample from a normal distribution of mean θ and known variance σ^2. Let the prior probability that $\theta = \theta_0$, the value on the null hypothesis, be c. Suppose that the remainder of the prior probability is distributed uniformly over some interval I containing θ_0. We shall deal with situations where $\bar{x}$, the arithmetic mean of the observations, and a minimal sufficient statistic, is well within the interval I. The posterior probability that $\theta = \theta_0$, in the light of the sample, can be evaluated; it is

$$\bar{c} = c \exp[-n(\bar{x}-\theta_0)^2/(2\sigma^2)]/K, \tag{1}$$

where

$$K = c \exp[-n(\bar{x}-\theta_0)^2/(2\sigma^2)] + (1-c)\int_I \exp[-n(\bar{x}-\theta)^2/(2\sigma^2)]\,d\theta,$$

by Bayes's theorem. In virtue of the assumption about $\bar{x}$ and I the integral can be evaluated as $\sigma\sqrt{(2\pi/n)}$.

Now suppose that the value of $\bar{x}$ is such that, on performing the usual significance test for the mean θ_0 of a normal distribution with known variance, the result is significant at the α percentage point. That is, $\bar{x} = \theta_0 + \lambda_\alpha \sigma/\sqrt{n}$, where λ_α is a number dependent on α only and can be found from tables of the normal distribution function. Inserting this value for $\bar{x}$ in (1) we have the following value for the posterior probability that $\theta = \theta_0$

$$\bar{c} = c\, e^{-\frac{1}{2}\lambda_\alpha^2}/\{c\, e^{-\frac{1}{2}\lambda_\alpha^2} + (1-c)\,\sigma\sqrt{(2\pi/n)}\}. \tag{2}$$

(Note that $\bar{x}-\theta_0$ tends to zero as n increases so that $\bar{x}$ will lie well within the interval I for sufficiently large n.) From (2) we see that as $n \to \infty$, $\bar{c} \to 1$. It follows that whatever the value of c, a value n can be found, dependent on c and α such that

(i) $\bar{x}$ is significantly different from θ_0 at the α % level;

(ii) the posterior probability that $\theta = \theta_0$ is $(100-\alpha)$ %.

This is the paradox. The usual interpretation of the first result is that there is good reason to believe $\theta \neq \theta_0$; and of the second, that there is good reason to believe $\theta = \theta_0$. The two interpretations are in direct conflict, and the conflict may apparently be made even stronger by remarking that the $(100-\alpha)$ % confidence and fiducial intervals for θ just exclude $\theta = \theta_0$. With $\alpha = 5$ we are 95 % confident that $\theta \neq \theta_0$, but have 95 % belief that $\theta = \theta_0$.

In commenting on this analysis, let us first consider the assumptions involved. Many significance tests involve situations in which the test criterion is asymptotically normally

distributed with known variance, as is $\bar{x}$ in the example, and therefore the sample considered is in no way unusual. The only assumption that will be questioned is the assignment of a prior distribution of any type, and, in particular, of the form chosen. A paradox will only have been generated if we can show there exist situations where (*a*) a prior distribution of this form is reasonable, and (*b*) a significance test of the 'tail-area' type is commonly used. Let us first consider the assignment of any prior probability. The argument for the use of prior probabilities has been put forward very cogently by Jeffreys (1948). His arguments have, to my mind, been reinforced by those of Ramsey (1931) and, more especially, Savage (1954). Savage's main contribution is as follows: he lays down certain axioms that a man should follow if he is to act in a 'rational' way, and defines a rational man to be a man who acts according to these axioms. The latter are quite mild in their form and would surely be agreed to by most statisticians. Savage then shows that a rational man must act as if he had a prior probability distribution and (if relevant) a utility function. It does not follow from this that any statistical inference need make overt mention of a prior distribution, but it does follow that no inference procedure should grossly contradict the existence of a prior distribution. (A mild contradiction may be allowable in the interests of simplicity.) Another way of looking at this result is to say that a probability distribution is a satisfactory measure of one's convictions about several hypotheses. For example, if to-day we say that our prior belief in one hypothesis is $\frac{1}{2}$ it will mean the same as saying to-morrow that our prior belief in a different hypothesis is $\frac{1}{2}$; just as a yard of material to-day measures the same as a yard of material to-morrow. If we are to use a significance level in a similar way, as Fisher (1956, p. 43) has suggested we can, and most statisticians do, we must establish a similar comparison property. 5% to-day must mean the same as 5% to-morrow. Our example, we claim, shows that it need not.

So much for the general question of introducing a prior distribution. We now consider the particular form used in deriving the paradox. We first note that the phenomenon would persist with almost any prior probability distribution that had a concentration on the null value and no concentrations elsewhere. For example, if there is an amount c at $\theta = \theta_0$ and the rest is distributed throughout I according to a density $p(\theta)$, where $\int_I p(\theta)\,d\theta = 1-c$, then if $p(\theta)$ is bounded it is easy to show, for example, by a steepest descent argument applied to the integral corresponding to that in (1), that $\bar{c}$ still tends to 1. It is sufficient that $p(\theta)$ does not tend to infinity too rapidly as θ tends to θ_0. It is, however, essential that the concentration on the null value exists, and it is this that has to be considered. Again Jeffreys (1948) has discussed the point. Briefly, one argument is that the singling out of the hypothesis $\theta = \theta_0$ to be tested is itself evidence that the value θ_0 is in some way special and is likely therefore to be true. We should like to give two examples where this seems unquestionably correct. The first is in genetics where θ is the linkage parameter between two genetic factors. If there is no linkage $\theta = \theta_0 = \frac{1}{4}$, and we are concerned with developing a test to determine if there is any evidence for linkage. Now in this situation there is a considerable amount of prior knowledge. For it is known that there is linkage if, and only if, the two genes lie on the same chromosome. Consequently if there are n chromosomes of approximately equal length, and if it seems reasonable to suppose that the gene is equally likely to be anywhere along the chromosomes' lengths, then it seems reasonable to suppose a prior probability of the order of $(n-1)/n$ that the value of θ is $\frac{1}{4}$. The particular numerical value of the prior probability is not so important here (though we note it is rather large)

as is the fact that $\theta = \frac{1}{4}$ is in a singular position and will arise for most positions of the genes. A second example arises in the telepathy experiments carried out by Soal & Bateman (1954), where, if no telepathic powers are present, the experiment has a success ratio of $\theta = \frac{1}{5}$, otherwise $\theta \neq \frac{1}{5}$. A significance test for telepathy therefore should assign to $\theta_0 = \frac{1}{5}$ a concentration of probability equal to one's prior belief that the subject has not got telepathic powers. This example is perhaps not as convincing as the genetical one because of the prejudices that exist in connexion with extra-sensory perception. My point in both these examples is that the value θ_0 is fundamentally different from any value of $\theta \neq \theta_0$, however near to θ_0 it might be. Unquestionably there exist situations (perhaps they are the more common) in which this is not so; where we are interested in testing the approximate validity of the null hypothesis, such as that the treatment has no (or very little) effect. This point has been discussed by Hodges & Lehmann (1954).

We now consider the paradox in these situations where the prior probability exists (by Savage's argument) and has a concentration on the null value. We first note that the expression of it in terms of fiducial or confidence limits used above is unjustified. The limits purport to be statements made about the value of θ in the light of the experimental result when initially nothing is known about or independent of knowledge of θ. The type of prior distribution used here (suggested by the practical circumstances of the problem) certainly does not correspond to ignorance about θ. Thus we should not be surprised at the disagreement. The paradox merely serves as a warning that the confidence or fiducial type of statement should only be used in those circumstances where one is truly ignorant about the parameter. We have argued that this is not so in the telepathy or genetical examples.

The conflict between statements of a significance level and statements based on Bayes's theorem remains. Now in our example we have taken situations in which the significance level is fixed because, as explained above, we wish to see whether its interpretation as a measure of lack of conviction about the null hypothesis does mean the same in different circumstances. The Bayesian probability is all right, by the arguments above; and since we now see that it varies strikingly with n for fixed significance level, in an extreme case producing a result in direct conflict with the significance level, the degree of conviction is not even approximately the same in two situations with equal significance levels. 5 % in to-day's small sample does not mean the same as 5 % in to-morrow's large one.

An alternative interpretation of the paradox was suggested to me by Prof. Barnard. The posterior probability $\bar{c}$, given by (2), may be written

$$\bar{c} = cf_n/\{cf_n + (1-c)\},$$

where

$$f_n = \sqrt{\left(\frac{n}{2\pi\sigma^2}\right)} e^{-\frac{1}{2}\lambda_\alpha^2},$$

the likelihood of θ_0 on the evidence of the sample. Clearly $f_n \to \infty$ as $n \to \infty$, λ_α fixed. Hence for fixed significance level the likelihood of the null hypothesis increases indefinitely with the sample size. This appears to me to demonstrate, without reference to prior probabilities, the unsoundness of the suggestion that significance tests depend on the disjunction: either a rare chance has occurred or the null hypothesis is false (Fisher, 1956, p. 39). For the chance considered in a significance test is the chance of the observed event and other more extreme ones. The chance of the observed event is measured by the likelihood function. These two chances behave quite differently. In fact, the paradox arises because the significance level

argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve. However, the above interpretation through the likelihood involves no mention of alternative hypotheses which seem basic to any approach to the problem.

The other approach to significance testing, due to Neyman & Pearson, does envisage the use of alternative hypotheses and hence appears to give a reason for using the tail area because this region is the best one in which to reject the null hypothesis at a specified level of significance. Therefore the occurrence of an observation in the region is an unusual event on the null hypothesis and less unusual on some alternative hypotheses. But the theory does not justify the practice of keeping the significance level fixed, nor does it take account of the fact that when the observation has been made we know, not that the point has fallen in the region of significance, but that it has fallen exactly on the edge, and the likelihoods under the null and alternative hypotheses seem the relevant quantities to compare.

The paradox is not, in essentials, new, although few statisticians are aware of it. The difference between the two approaches has been noted before by Jeffreys (see, in particular, 1948, Appendix), who is the originator of significance tests based on Bayes's theorem and a concentration of prior probability on the null value. But Jeffreys is concerned to emphasize the similarity between his tests and those due to Fisher and the discrepancies are not emphasized. The same phenomenon was noticed by Lindley (1953) in decision theory studies, and some computations by Prof. Pearson in the discussion to that paper emphasized how the significance level would have to change with the sample size, if the losses and prior probabilities were kept fixed. (The discussion based only on the latter quantities is mathematically equivalent to one in decision theory language with zero-one losses.) The present note considers the situation where the significance level is fixed and the variation in posterior probability is evaluated, rather than the other way round.

The concept of a significance level has been used very successfully in practical problems of inference. One might now ask how this has come about. The answer has already been given by Jeffreys in the appendix already cited. Essentially it is because $\bar{c}$, as given by (2), tends to unity very slowly and, for moderate values of n, $\bar{c}$ may be less than c at a prescribed significance level and the two concepts be in reasonable agreement. Let

$$A = c\,e^{-\frac{1}{2}\lambda_\alpha^2}/(1-c)\sqrt{(2\pi)},$$

then

$$\bar{c} = A/(A+\sigma/\sqrt{n}), \tag{3}$$

and $\bar{c} \to 0$ as $\sigma/\sqrt{n} \to \infty$. Hence in a small experiment, significance at 5 % may give very strong reasons to doubt the null hypothesis. A numerical example is informative. Suppose we take $c = \frac{1}{2}$ and use a two-sided test at 5 % significance so that $\lambda_\alpha = 1{\cdot}96$; then $A = 0{\cdot}0584$ and the table gives the value of $\bar{c}$ for different values of $t = n/\sigma^2$. If $\sigma = 1$, $t = n$, and we see that for small samples ($n \leqslant 10$) the probability of θ_0 has decreased appreciably from its initial value of $\frac{1}{2}$, giving cause to doubt the validity of the null hypothesis. For medium samples ($10 < n < 100$) the probability has only decreased a little, so that although we are not as confident as we were initially about the null hypothesis, our doubts are not great. By the time n has reached a value about 300 $\bar{c}$ is equal to c; the experiment, despite its 5 % significance, has not altered our belief in the null hypothesis at all. To reach the strong contrast put forward in the paradox it would be necessary to take n about 10,000. Of course if σ is smaller then smaller samples will suffice. For example, if we apply these numerical values to the Soal & Bateman problem (i.e. use the normal approximation to the binomial) we have $\sigma^2 = \frac{1}{5}.\frac{4}{5} = 0{\cdot}16$, and a sample of size about 48 has $\bar{c}$ equal to the

original value of $\frac{1}{2}$. An experiment of this type with a run of forty-eight trials which is significant at 5% would not alter our views on telepathy if initially we had an open mind on the problem. The normal approximation is not adequate for samples as low as 10, but it is clear that only such small ones would increase our prior belief at all noticeably. An experiment of 1600 trials would raise our belief that telepathy did not exist to 95%; quite a moderate size in comparison with the 37,100 trials carried out with Mrs Stewart. The reader may be interested to know that with $c = \frac{1}{2}$ the posterior probability of the null hypothesis $\theta = \frac{1}{5}$ in the light of the experiments with Mrs Stewart (9410 successes) is of the order of 10^{-140}. The evidence for Mrs Stewart's telepathic powers is rather strong.

t	$\bar{c}$	t	$\bar{c}$
1	0·055	600	0·589
2	0·076	800	0·623
3	0·092	1,000	0·649
4	0·105	2,000	0·723
5	0·116	4,000	0·787
10	0·156	6,000	0·819
20	0·207	8,000	0·839
40	0·270	10,000	0·854
60	0·312	20,000	0·892
80	0·343	40,000	0·921
100	0·369	60,000	0·935
200	0·453	80,000	0·943
300	0·503	100,000	0·949
400	0·539	∞	1·000

An apparent advantage of the significance level statement is that it does provide some sort of assessment of the truth of the null hypothesis using only the evidence provided by the experiment. It is, in effect, a convenient (though possibly misleading) summary of what the experimental result has to say about the null hypothesis. A similar assessment is available in a Bayesian analysis through the likelihood function. In the situation considered here the function is proportional to

$$\sqrt{\left(\frac{n}{2\pi}\right)} \exp\{-\tfrac{1}{2}n(\bar{x}-\theta)^2\},$$

regarded as a function of θ. This, unlike the single number expressing the significance level, is a function and is therefore more difficult to understand. A reduction to a numerical value is possible provided the assessment of prior probabilities conditional on $\theta \neq \theta_0$ is made. For example, if θ is uniform in the interval I in these circumstances, then

$$\exp[-n(\bar{x}-\theta_0)^2/2\sigma^2] \Big/ \int_I \exp[-n(\bar{x}-\theta)^2/2\sigma^2]\,d\theta = \sqrt{\left(\frac{n}{2\pi}\right)} \exp\left\{-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\right\}$$

is the quantity by which the prior odds, $c/(1-c)$, in favour of θ_0 must be multiplied in order to obtain the posterior odds, $\bar{c}/(1-\bar{c})$. This single value, or its logarithm, might be an acceptable substitute for the significance level. It is numerically equal to Jeffreys's K, since he supposes $c = \frac{1}{2}$.

The paradox serves to explain one puzzling feature of tests based on Bayes's theorem. Suppose the experimenter has continued sampling randomly until he has reached a result which is, using a fixed-sample size significance test, significant at some prescribed significance level α. That is, he has taken a sample $(x_1, x_2, \ldots, x_n)$ such that $\bar{x} = \theta_0 + \lambda_\alpha \sigma/\sqrt{n}$.

It is easy to show, by the law of the iterated logarithm, that this will happen with probability one whatever the value of θ. Then the experimenter has, of course, cheated if he quotes his result as being significant at $\alpha\%$, though, if the distribution theory were known, a valid significance test could be made. But it would not be that appropriate to a sample of fixed size n. On the other hand, it is easy to see that the likelihood of the observations $(x_1, x_2, \ldots, x_n)$ does not depend on the particular sequential stopping rule used and is, therefore, equal to the likelihood the experimenter would have obtained if the same sample had been reached by taking a sample of fixed size n. It follows that any significance test based on Bayes's theorem does not depend on the sequential stopping rule used, at least amongst a wide class of such rules. In the extreme case the experimenter can go on sampling until he has reached the significance level α, and yet the fact that he did so is irrelevant to a Bayesian. In telepathy this is known as 'optional stopping': stopping when the results look striking; striking, that is, on a significance level criterion. The explanation is now clear. If $\theta \neq \theta_0$ the optional stopper will reach his desired point for small n and $\bar{c} < c$. On the other hand, if $\theta = \theta_0$ the value of n will be larger and $\bar{c} > c$. (These are average results, of course, naturally sometimes mistakes will be made.) The value of $\bar{c}$ is just what one would expect in the two cases and we see that the Bayesian will not on the average be in error in ignoring the stopping rule. It should now be possible to give a reliable assessment of those results in telepathy which have had objections raised against them on the grounds of optional stopping.

I am much indebted to Profs. Pearson and Barnard for helpful comments on the first draft of this paper.

REFERENCES

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.
HODGES, J. L. & LEHMANN, E. L. (1954). *J. R. Statist. Soc.* B, **16**, 261–8.
JEFFREYS, H. (1948). *Theory of Probability*, 2nd ed. Oxford: Clarendon Press.
LINDLEY, D. V. (1953). *J. R. Statist. Soc.* B, **15**, 30–76.
RAMSEY, F. P. (1931). *The Foundations of Mathematics.* London: Routledge and Kegan Paul.
SAVAGE, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.
SOAL, S. G. & BATEMAN, F. (1954). *Modern Experiments in Telepathy.* London: Faber and Faber.