SOME OBSERVATIONS ON LAPLACE'S RULE OF SUCCESSION AND PERKS'S RULE OF INDIFFERENCE

By JOHN E. FREUND

Associate Professor of Mathematics, Alfred University, New York

THE controversial problem of prior probabilities, inverse inference, and the rule of Bayes-Laplace has recently been brought back into the limelight through Wilfred Perks's interesting attempt to formulate a new indifference rule for the prior probabilities in the theory of inverse probability.* The purpose of this paper is to subject this new theory to a critical appraisal from the point of view of the practical statistician. Perks describes the desirable features of his method by pointing to certain invariance properties of his prior probabilities, their analogy to the Lorentz transformation in the theory of relativity, the (according to him) 'naturalness' of angles being uniformly distributed, and other supposedly 'reasonable' features. It is our contention that the comparison of different sets of assumptions and their resulting mathematical systems cannot be based on such 'mathematical niceties'. The physicist does not use Riemannian geometry in preference to Euclidean geometry because he likes its peculiar mathematical properties; he uses it because it supplies him with a more suitable framework to study the phenomena of nature. Similarly, if we are interested in developing a mathematical theory which is to furnish us with formulae for estimating probabilities, the merits of whatever assumptions we happen to employ can be judged only on the basis of the 'edible fruit', i.e. on the basis of the most relevant features of the estimators which are derived from the assumptions.

We shall not, therefore, dwell upon the particular assumptions which were made by Perks, but we shall compare the estimator of p which he obtained with the corresponding maximum-likelihood estimator and Laplace's famous rule of succession. The general problem with which we are concerned is to estimate the probability of obtaining a 'success' in the next trial if m 'successes' were observed in n previous trials with the customary assumptions of independence and constant probability. In other words, we are trying to estimate the parameter p of a binomial distribution. The maximum-likelihood estimator of p is given by the relative frequency

$$f' = \frac{m}{n}, \tag{1}$$

Laplace's rule of succession which is based on a uniform prior distribution supplies us with the estimator

$$f'' = \frac{m+1}{n+2}, \tag{2}$$

while Perks's new indifference rule produces the estimator

$$f''' = \frac{m+\frac{1}{2}}{n+1}. \tag{3}$$

* WILFRED PERKS, Some observations on inverse probability including a new indifference rule, *J.I.A.* LXXIII, 285.

The relative merits of several estimators of one single parameter are usually evaluated either in terms of which estimator hits the exact value of the parameter 'precisely on the nose' with the highest frequency, or in terms of which estimator involves on the average the smallest error, i.e. in terms of efficiency. Since the first of these two criteria is seldom of particular value, with the notable exception of the circus side-show performer who guesses a person's age and loses unless he guesses it correctly, we shall compare the above three estimators in terms of their sampling errors which are defined as

$$s_f^2 = \mathrm{E}(f-p)^2. \tag{4}$$

This expected value is given by the summation

$$\sum_{m=0}^{n} (f-p)^2 \binom{m}{n} p^m (1-p)^{n-m}.$$

The expression differs slightly from the customary standard error because we are dealing with estimators which need not necessarily be unbiased, in which case (4) is a more reasonable evaluation of the sampling error.

Since all three of the above estimators are of the form

$$f = \frac{m+bk}{n+b}, \tag{5}$$

we shall first evaluate the sampling error of f and then treat f', f'' and f''' as special cases with the proper choice of the numerical values of b and k.

If we use the first two moments of the binomial distribution, it can easily be shown that

$$s_f^2 = \frac{p^2(b^2-n) + p(n-2kb^2) + b^2k^2}{(n+b)^2}, \tag{6}$$

and if we substitute suitable values for b and k we obtain

$$s_{f'}^2 = \frac{p(1-p)}{n}, \tag{7}$$

$$s_{f''}^2 = \frac{p(1-p)(n-4)+1}{(n+2)^2}, \tag{8}$$

$$s_{f'''}^2 = \frac{p(1-p)(n-1)+\frac{1}{4}}{(n+1)^2}. \tag{9}$$

A most unfortunate property of all of these sampling errors is that they are functions of the parameters which we are trying to estimate. This is not so if, for example, we estimate the mean of a normal population. If we may borrow from the language of Neyman and Pearson, we could say that the sample mean is a uniformly more efficient estimator of the mean of a normal population than the sample median, since its sampling error is smaller regardless of the actual value of the population parameter.

It can easily be seen from (7), (8) and (9) that no one of the above three estimators is uniformly more efficient than the others and, as a matter of fact, it can easily be shown that no uniformly most efficient estimator exists.

[Perks also demonstrated that if he were to maximize the posterior probability his rule of indifference would produce the estimator

$$f'''' = \frac{m-\frac{1}{2}}{n-1}. \tag{3a}$$

[If $b=-1$ and $k=\frac{1}{2}$ are substituted in (6) we obtain

$$s^2_{f''''}=\frac{p(1-p)(n-1)+\frac{1}{4}}{(n-1)^2}, \tag{9a}$$

and a comparison of (9) and (9*a*) shows that $s^2_{f''''}$ is less than $s^2_{f'''}$ regardless of p. This proves that f''' is uniformly more efficient than f''''.]

Since we can make no such statement concerning f', f'' and f''', let us now investigate the ranges of the values of p for which each of these three estimators is (relative to the two others) the most efficient.

A direct comparison of (7) and (8) shows that f'' is more efficient than f' if p lies in the interval

$$\text{from}\quad p=\frac{1}{2}-\frac{1}{2}\sqrt{\frac{(n+1)}{(2n+1)}}\quad\text{to}\quad p=\frac{1}{2}+\frac{1}{2}\sqrt{\frac{(n+1)}{(2n+1)}}. \tag{10}$$

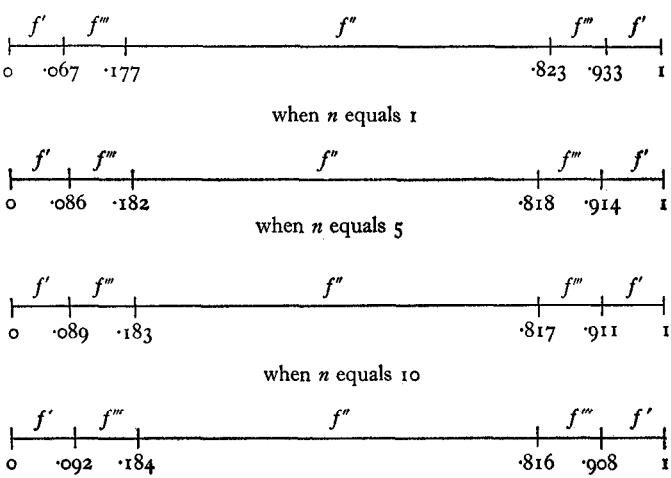
Similarly, it follows from (7) and (9) that f''' is more efficient than f' if p lies in the interval

$$\text{from}\quad p=\frac{1}{2}-\frac{1}{2}\sqrt{\frac{(2n+1)}{(3n+1)}}\quad\text{to}\quad p=\frac{1}{2}+\frac{1}{2}\sqrt{\frac{(2n+1)}{(3n+1)}}. \tag{11}$$

Finally, f'' is more efficient than f''' if p lies in the interval

$$\text{from}\quad p=\frac{1}{2}-\frac{1}{2}\sqrt{\frac{(2n+3)}{(5n+7)}}\quad\text{to}\quad p=\frac{1}{2}+\frac{1}{2}\sqrt{\frac{(2n+3)}{(5n+7)}}. \tag{12}$$

In order to demonstrate the intervals of p on which f', f'' and f''' are, relatively speaking, most efficient, the intervals are computed for $n=1$, $n=5$, $n=10$, and when n approaches infinity. The particular estimator which is most efficient is indicated for each interval of the following diagram:



It is tempting to conclude from this diagram that, since f'' is the most efficient estimator over a considerably larger interval of values of p than either of the other two estimators, it is, consequently, the most desirable estimator. This is,

of course, not a permissible inference unless we make further assumptions about the prior distribution of p. If the probabilities which a statistician estimates over a considerable period of time vary pretty uniformly over the interval from 0 to 1, he would be justified in preferring f'' to either f' or f'''. If, on the other hand, most of the probabilities which he estimates are very small, he would do better by using f'. As Perks's paper seems to indicate, this is the case in actuarial statistics where our main concern is to estimate annual mortality rates. Since such rates rise gradually from about ·0015 at age 10 to ·085 at age 75, f' is always better than the other two. This is true, at least, for this range of ages which is by far the most important in practical applications.

It is evident from this discussion that no absolute comparison of the three estimators is possible without the introduction of further assumptions concerning the values of p which most frequently occur in practice. Such assumptions could be avoided if we would be satisfied with a comparison in the 'minimax' sense. The three sampling errors which were given in (7), (8), and (9) are each at a maximum when $p=\frac{1}{2}$. If we now substitute this value, we obtain

$$\max s^2_{f'} = \frac{1}{4n}, \tag{13}$$

$$\max s^2_{f''} = \frac{n}{4(n+2)^2}, \tag{14}$$

$$\max s^2_{f'''} = \frac{n}{4(n+1)^2}. \tag{15}$$

A comparison of these three maximum values of the s^2 shows immediately that $s^2_{f''}$ has the smallest maximum value.*

It has been our purpose to demonstrate that in spite of the mathematical niceties which underlie Perks's formula there appears to be no reason why it should be preferred to either f' or f''. As a matter of fact it is apparent from our discussion that f''' is more efficient than f' and f'' only for values of p which fall on a very special and extremely small interval. The choice between f', f'', f''' and possibly other estimators seems to us to be essentially a matter of taste unless, of course, we do have some information concerning the prior probabilities. If the prior probabilities can be estimated, as in certain problems in casualty insurance, the most suitable estimator can be derived directly from the rule of Bayes-Laplace. In that case we have to concern ourselves with the prior probabilities of the prior probabilities pushing, thus, the basic problem of statistical inference to a lower level.†

* It has recently been shown by J. L. HODGES and E. L. LEHMANN, 'Some problems in minimax point estimation', *Ann. Math. Statist.* XXI, no. 2 (1950), that the maximum value of s^2 is least for the estimator

$$\frac{\sqrt{n}}{\sqrt{n}+1}\frac{m}{n}+\frac{1}{\sqrt{n}+1}\frac{1}{2}.$$

This estimator is, however, more efficient than the estimators which we have discussed only when p is very close to $\frac{1}{2}$. The length of the interval where it is most efficient approaches 0 with increasing n.

† If we interpret the prior probabilities of a parameter θ in the sense of R. VON MISES, 'On the correct use of Bayes's formula', *Ann. Math. Statist.* XIII, 1942, a discrete set of such prior probabilities $f(\theta)$ can itself be considered as a set of population parameters which are to be estimated by means of some standard technique, e.g. maximum likelihood.

It is interesting that the various estimators which we have discussed may be considered as weighted estimators by writing

$$f=\frac{m+bk}{n+b}=\frac{n}{(n+b)}\frac{m}{n}+\frac{b}{(n+b)}k. \tag{16}$$

In this formula the weight $\frac{n}{n+b}$ is given to the relative frequency m/n, while the remaining weight, namely, $\frac{b}{n+b}$, is given to the constant k. These weights are such that when n is small most of the weight goes to the constant k, but it shifts towards the relative frequency with increasing n. We have, therefore, a weighting procedure which combines direct information (the relative frequency m/n) with collateral information (the constant k).

Formula (16) is, in principle, identical with the credibility formulae which have been used for many years by casualty actuaries in the United States. Such formulae weight the direct experience of a risk and some collateral information concerning a wider class of risks to which the particular risk has been assigned. If we have relatively little direct knowledge concerning an individual risk, then most of the weight is given to a constant representing the entire classification. The more direct experience we have, however, the more weight is shifted towards the direct knowledge. This is precisely what happens also in formula (16).

If we write our three estimators in this form we obtain

$$f'=\frac{m}{n}, \tag{17}$$

$$f''=\frac{n}{n+2}\frac{m}{n}+\frac{2}{n+2}\frac{1}{2}, \tag{18}$$

$$f'''=\frac{n}{n+1}\frac{m}{n}+\frac{1}{n+1}\frac{1}{2}, \tag{19}$$

and it becomes apparent that, while f' gives the entire weight to the relative frequency, both f'' and f''' divide the weight between the relative frequency m/n and the constant $\frac{1}{2}$. The only distinction between f'' and f''' is that the weight is shifted faster towards the relative frequency in the formula which represents f'''. Both f'' and f''' are subject to the same criticism that some of the weight is given to the constant $\frac{1}{2}$. Possibly this value might be justified in the sense of a minimax solution. In the absence of any direct knowledge the choice of the value $\frac{1}{2}$ involves the minimum risk, i.e. it is such that it minimizes the maximum error. If a suitable value of k is known, however, a formula like that which is given in (16) would provide us with a much more desirable estimator.

The purpose of this paper has not been to develop any new methods of estimation but rather to shed some light on the problem of the estimation of probabilities by discussing some of its practical aspects which are usually not mentioned in the more profound analyses of the foundations of statistical inference.

[Professor Freund sent a copy of the foregoing article to Mr Perks who has submitted the following comments.—Eds. *J.I.A.*]

Professor Freund states that the purpose of his article is 'to subject this new theory' (i.e. the invariant indifference rule for prior probabilities) 'to a critical appraisal from the point of view of the practical statistician'. In fact, his article deals with only a particular aspect of the subject; it is largely confined to a comparison of the theoretical merits of various point-estimates of the binomial parameter—the maximum-likelihood estimate m/n, the Bayes-Laplace estimate $(m+1)/(n+2)$, the invariant-rule mean value estimate $(m+\frac{1}{2})/(n+1)$ and the invariant-rule maximum posterior-probability estimate $(m-\frac{1}{2})/(n-1)$. For this comparison he chooses as his criterion of merit the arbitrary measure of statistical 'efficiency' used by a particular school (or schools?) of mathematical statisticians. I am not sure that they themselves would nowadays claim that the second moment measured from the (unknown) parameter value is the best measure of 'efficiency' in a case, such as the one under discussion, where the distribution of the estimate does not—except in the limit—follow the normal curve and is, in fact, skew. At any rate, the second moment, as a measure of efficiency, has a fundamental significance only for estimates which are distributed normally. There is a close analogy in the preference by Beard, myself and others for the wD test* over the χ^2 test for testing mortality table graduations. However, the inappropriateness of Professor Freund's criterion in practice is apparent from the fact that it selects the maximum-likelihood estimate in those extreme cases of 100% successes and 100% failures for which no reasonable person would accept its verdict of certainty and impossibility respectively.

I find it difficult to think that the practical man—whether a 'practical statistician' or an actuary—cares in the slightest degree for these minute differences in point-estimates. Has he not already dismissed the whole thing as hair-splitting? For my part, I expressly stated in my paper that it had no practical significance whatever. I was concerned solely with removing the long-standing *logical* objections to inverse probability, i.e. the alleged inconsistency arising on transformation of the parameter, and the absence of a frequency interpretation. The former objection was removed by the invariant indifference rule and the latter by stressing that all probability (of whatever school) contained an essential non-frequency element however much it might be concealed by partisan expositors. Subsequent work has not disturbed my claims on the former question and has, I think, tended to confirm them on the latter. Professor Freund ignores all this except that he refers to certain 'desirable features of the method' as if they were objectives rather than consequences of the theory—this is to put the cart before the horse. These features are indeed some of the 'edible fruit' by means of which the theory may appropriately be judged.

W. P.

* See p. 391.