WHICH COMES FIRST, PROBABILITY OR STATISTICS?

By I. J. GOOD, M.A., Ph.D.

THE title of this note was selected so as to provide an excuse for discussing some rather general matters. Let us first consider the question of which of probability and statistics came first historically. This question is like the one about eggs and chickens. The question whether eggs or chickens came first could in principle be given a meaning by using arbitrarily precise definitions of eggs and chickens, and even then probably nobody would be able to answer the question. The question whether probability or statistics came first is not as bad as the one about eggs and chickens, but it still depends on arbitrary definitions.

Some of the ideas of probability and statistics must be very old. Cicero described probability as 'the guide to life', and insurance was practised by the ancient Romans. (A table of life expectancies was constructed by Domitius Ulpianus.) Even animals behave as if they accepted some of the principles of scientific induction, so that inductive behaviour can be said to date back to a time before there were any human beings.

A substantial use of mathematics in the theory of probability was apparently not made until 1654 in correspondence between Pascal and Fermat, published in 1679. A much more convincing use of mathematics was made in Huygens's dissertation, published in 1657. Slight anticipations of the mathematical theory are mentioned in Todhunter's *History of the Mathematical Theory of Probability*. For example, a commentary of 1477 on Dante's *Purgatorio* refers to the probability of various throws with three dice. Galileo made similar calculations before 1642, published in 1718; apparently there were delays in publication in those days! Both Pascal and Galileo were stimulated by questions put to them by gamblers, who had accumulated statistical data for which they requested explanations. Cardan, himself a gambler, made similar calculations perhaps a hundred years earlier even than Galileo. As pointed out by Miss F. N. David (*Biometrika*, **42** (1955), 1–15), Cardan had explicitly formulated the method of basing the calculation of probabilities on the abstraction of equally probable cases. Clearly Cardan and other gamblers knew that the limiting relative frequencies of successes ought to be equal to the proportion of equally probable ways of getting a success, otherwise they would not have asked for explanations. In other words, even before the mathematical theory of probability was founded, it seems that many people must have had a rough idea of the strong law of large numbers, although this law was not properly deduced from the usual axioms until 1917 (by Cantelli, generalizing work of Borel and Hausdorff).

It is at least conventional then to say that the mathematical theory of probability was founded in the middle of the seventeenth century.

Fermat and Pascal, and the other writers mentioned, started the mathematical theory of probability in order to explain the results of some statistics obtained experimentally, so that it could be contended that statistics came first. But since we have dated probability by mathematical probability, the reasonable question is 'When did *mathematical* statistics start?' It seems fair to say that it started with either (*a*) de Moivre's work on life annuities (1718) (the earlier

work of Halley (1693) was mathematically trivial) or (*b*) the 'theory of errors' or combination of observations. The latter subject was developed greatly by Laplace and Gauss, about 1819–27, but it had its origins in work by Thomas Simpson (1757), Lagrange (about 1770) and Daniel Bernoulli (1777). Daniel Bernoulli used the method of maximum likelihood, but did not discover the method of least squares because he did not use the normal law of errors.

Apparently, then, the mathematical theory of statistics started at least sixty years later than that of probability. In fact, mathematical statistics is largely based on mathematical probability, so that mathematical probability is practically bound to take both historical and logical precedence. A possible loophole in this statement is that statistics is not entirely predictive and inferential; some of it is concerned merely with the reduction of a great deal of data to manageable form. If there are enough data for the theory of probability to be in effect ignored, then the statistical problems tend to become mathematically rather trivial.

The question now arises whether statistics requires any theory or technique outside the theory of probability. It will be more interesting to consider the slightly different question whether statistics requires anything outside the 'theory of rational behaviour', by which is meant the theory of probability and utility.

In order to give statistics the best chance of belonging to the theory of rational behaviour we may take this theory in its most general sense, i.e. with a subjectivistic or multisubjectivistic interpretation, both for the probabilities and the utilities. Such a theory is more general than a 'necessary' theory in which probabilities measure objective rational degrees of belief called credibilities, usually assumed to be precise and sometimes determinable. (Necessary theories of probability were adopted by H. Jeffreys and R. Carnap, and also by J. M. Keynes, who, however, retracted in his biography of F. P. Ramsey, *Essays in Biography* (1933).) Moreover, a necessary theory is more general than a frequentist theory, in which probabilities are all limiting frequencies.

A subjectivistic theory breaks up into three parts:*

(1) Axioms
(2) Rules of application } Theory } Technique.
(3) Suggestions

(In fact, any applied theory of probability, or applied scientific theory in general, can be broken up into three such parts.) The axioms give rise to an abstract theory that can be applied only if rules of application are given. The axioms and rules together make up the theory of probability, and then there is an indefinite number of 'suggestions' (discussed below) that are better regarded as belonging to the 'technique' of probability rather than to the theory proper. The axioms can be expressed in terms of propositions, for which we use the symbols E, F, G and H. (Some mathematicians prefer to base the axioms on sets instead of on propositions, but this seems to me to make the axioms more abstract without gain of generality.) Symbols of the form $P(E \mid F)$, $U(A \mid H)$ are introduced and are read 'the probability of E given F' and 'the utility of A given H', where A is an act and H is a proposition that is to be interpreted later as a description of an assumed state of the world.

* Cf. *Probability and the Weighing of Evidence* (1950, written, for the most part, in 1946). Similar views had previously been expressed by F. P. Ramsey and Bruno de Finetti.

$P(E \mid F)$, for all propositions E and F, is a real number between 0 and 1, while $U(A \mid H)$ is also a real number. (These assumptions are made for the sake of simplicity. B. O. Koopman (1940), inspired by J. M. Keynes, worked out a more general theory, but without utilities, in which probabilities are non-numerical and only partially ordered, even inside the abstract theory.) A typical axiom is

$$P(E \text{ and } F \mid G) = P(E \mid G).P(F \mid E \text{ and } G).$$

A typical rule of application is that a judgment

$$P'(E \mid F) > P'(G \mid H), \tag{1}$$

meaning that the degree of belief in E given F would be greater than that in G given H, permits the inequality

$$P(E \mid F) > P(G \mid H) \tag{2}$$

to be used in the abstract theory; and conversely that (2) implies the 'discernment' (1). The principle of rational behaviour is one of the rules of application. It is the recommendation to maximize expected utility.

The purpose of the theory is to increase the size of a body of beliefs or preferences between acts and to detect inconsistencies in it, and thereby to bring some degree of objectivity into our beliefs and decisions. The purpose of a subjectivistic theory is to increase objectivity.

The definitions of probability and utility are implicit in the theory as a whole, i.e. in the axioms and rules of application. Rigorous explicit definitions are not given.

Suggestions are 'vague rules'. For example,

(i) the theorems of probability, such as the laws of large numbers, can be used to help the judgment;

(ii) if an inconsistency is found, it should be resolved by means of honest and detached judgment;

(iii) occasional inconsistencies may be tolerated, such as when we regard the probability of a mathematical theorem as neither 0 nor 1 (cf. *Probability and the Weighing of Evidence*, p. 49, and also G. Polya's work on plausible inference);

(iv) numerical probabilities can be introduced either by imagining idealized games of chance, or (equally idealized) infinite sequences of trials performed under essentially the same circumstances;

(v) lower bounds for very small probabilities can be estimated by the 'device of imaginary results' in which one imagines a sequence of successful trials sufficient to bring the probability up above 1/2, and then applies Bayes's theorem in reverse to get a bound on the initial probability (cf. *Probability and the Weighing of Evidence*, pp. 35 and 70, for a fuller explanation of this device).

We may now ask why the theory of rational behaviour is not sufficient for statistics. The answer is that in statistics we usually aim at precise statements for summarizing evidence. We cannot always arrive at precise statements in the theory of rational behaviour because in this theory we deal primarily with inequalities, not with equations. To overlook this fact is like imagining that lengths can be found to an infinite number of places of decimals, a mistake that is not made in an engineering specification in which tolerances are specified.

It is possible, however, to regard statistics as belonging to the *technique* of rational behaviour. The principles of statistics are mostly suggestions in the technique of rational behaviour or of probability.

No principle of statistics is uncontroversial when expressed too precisely, unless it is already a part of the theory of rational behaviour. For if it were uncontroversial and precise, it would have been adjoined to the axioms of the subject.

The theories of probability and of rational behaviour are extensions of ordinary logic and must take logical precedence over statistics. This precedence of probability and utility is sometimes overlooked when people take the apparent precision of statistical principles too seriously. They are then liable to forget to check the consistency of each application with their own honest judgments, and they think they are thereby exhibiting a noble objectivity.

It is often said that the foundations of probability are controversial. But the controversy is perhaps illusory and is concerned only with matters of terminology, such as whether the word 'probability' should be given a very wide meaning or no meaning at all. The controversies in statistics are more real unless it is admitted that the principles of statistics are usually imprecise. It is difficult for statisticians to make such an admission when one of the main aims of statistics is the avoidance of vagueness.

Let us consider some examples of statistical principles. For each of them we shall run into trouble by regarding them as golden rules leading to precise probability statements or decisions.

(i) *Maximum likelihood.* I recently won a one-cent bet by guessing the name of the last entry in a dictionary of 50,000 American scientists. (It was Zygmund.) The maximum-likelihood estimate of the number of names of American scientists known to me, on this evidence, is 50,000—clearly an unreasonable estimate. Fisher would recommend that the principle of maximum likelihood should be used with common sense. Another way of saying the same thing would be that initial probabilities and utilities should be taken into account. (For an example where maximum likelihood gets into trouble even for large samples see Lindley, *J.R. Statist. Soc.*, 1947.)

(ii) *Tail-area probabilities.* One of the earliest attempts to avoid the use of more than the minimum of judgment was the use of tail-area probabilities (the so-much-or-more method). A typical example is the use of χ^2 by Karl Pearson (1900). An earlier use was by Laplace (1773) in a memoir on the inclination of the orbits of comets. There was an earlier, but rather trivial, example by Arbuthnot (1712). But presumably gamblers must have used the method in a rough and ready way, even before 1654, for deciding whether to draw swords on their opponents for cheating.

To prove that the use of tail-area probabilities as the final summary of statistical evidence is controversial, it is sufficient to refer to a paper by Neyman & E. S. Pearson (1928), in which it was emphasized that likelihoods on non-null hypotheses are relevant as well as those on the null hypothesis. It is possible to regard this emphasis as constituting a slight swing back to the Bayes–Laplace philosophy. (An example to show that the probability distribution of a statistic on the null hypothesis is not enough for determining the choice of which statistic to use is that the reciprocal of Student's t has the same distribution as t itself when the sample is of size 2.)

(iii) *Large-sample theory, or asymptotic properties of statistics.* A good

deal of modern statistical theory is concerned with the asymptotic properties of statistics. One controversial question is how large samples have to be in order to make these asymptotic properties relevant.

(iv) *The likelihood-ratio method.* In this method a statistic is chosen that is equal to the ratio of maximum likelihoods among the class of simple statistical hypotheses being tested and among the class of all simple statistical hypotheses entertained. Though intuitively appealing and having desirable large-sample properties, a small-sample example was produced by Stein in which the method leads to absurd conclusions (see, for example, J. Neyman, *Lectures and Conferences on Mathematical Statistics and Probability* (1952)).

(v) *Unbiased statistics.* Unbiased statistics can take values outside the range of what is possible. For example, if a multinomial distribution has category chances $p_1, p_2, \ldots, p_m$, and if in a sample of size N the frequencies of the m classes are $n_1, n_2, \ldots, n_m$, then an unbiased estimate of Σp_i^2 is

$$\Sigma n_i(n_i - 1)/N(N-1).$$

This estimate would vanish if each n_i were either 0 or 1, but the minimum possible value of the population parameter is $1/m$. It is tempting to replace by $1/m$ those values of the statistic that turn out to be less than $1/m$. Some statisticians would do this without noticing that they were now using a biased statistic.

It is sometimes argued that unbiased statistics have an advantage if it is intended to average over a number of experiments. Two questions then arise: (*a*) How many experiments? (*b*) Would a modified Bayes–Laplace philosophy do just as well if not better? (By the 'modified Bayes–Laplace philosophy' we mean the philosophy described in the present note. It differs from the ordinary Bayes–Laplace philosophy in that it leaves room for individual judgment instead of assuming uniform initial distributions, i.e. Bayes postulates.) No modified Bayes–Laplace estimate can lie outside the possible range of values of the population parameter. Applications of the modified Bayes–Laplace philosophy do not yet belong to orthodox statistics. They are not intended to lead to precise results.

(vi) *Deciding on significance tests before taking a sample.* In elementary text-books the advice is often given to decide on one's tests of significance before taking a sample. This may be good advice to those whose judgment you do not trust. Or a statistician may use the principle for himself as a precaution against wishful thinking, or as a guarantee against accusations of prejudice rather than judgment. But consider the following example. A sample of 100 readings is taken from some distribution for which the null hypothesis is that the readings are independently distributed with a normal distribution of zero mean and unit variance. It is decided in advance of sampling to divide this normal distribution up into ten equal areas, and to apply the χ^2 test to the ten-category equiprobable multinomial distribution of frequencies with which the readings fall into the ten areas. This would appear to be a very reasonable statistic. But what if it leads to a non-significant result even though one of the 100 readings was 20 standard deviations above the mean?

(vii) *Confidence intervals.* (Developed mainly by Neyman & Pearson (1930–3). Suggested by E. G. Wilson, *J. Amer. Statist. Ass.* (1927). I am indebted to Prof. S. S. Wilks for this last reference.) One of the intentions of using confidence intervals and regions is to protect the reputation of the statistician by being right in a certain proportion of cases in the long run.

Unfortunately, it sometimes leads to such absurd statements, that if one of them were made there would not be a long run. One objection, similar to the one above concerning unbiased estimates, was given by M. G. Kendall, *Biometrika*, **36** (1949), 101–16. For others see the discussion on H. Daniels, 'The theory of position-finding,' *J. R. Statist. Soc.* B, **13** (1951). A further objection is admitted in Neyman's book, cited under heading (iv), namely, that it can lead to absurdly long confidence intervals. Stein introduced a sequential sampling procedure to overcome this last objection, but it can lead to absurdly large samples.

A statistician can arrange to make confidence pronouncements that are correct in at least 95% of cases in the long run (if there is a long run). But if his customer decides to separate off the pronouncements that relate to a subclass of the possible experimental results (such as those in which a random variable is large), then it is no longer true that 95% of the subclass will be correct in general. In fact the judgment that the random variable is large is an indirect statement about the initial probability distribution, and it will imply that for this subclass the proportion of correct confidence interval statements will probably fall below 95%. This argument shows what is perhaps the main reason why the confidence method is a confidence trick, at least if used too dogmatically.

(viii) *Fiducial distributions.* The use of fiducial distributions in statistical inference is controversial if only because these distributions need not be unique. (See J. G. Mauldon, *J. R. Statist. Soc.* B, **17** (1955), 79–95. There is similar unpublished work by J. W. Tukey.)

(ix) *Errors of the first and second kinds.* The notion of the minimization of sampling costs for a given consumer's risk was used by Dodge & Romig, *Bell Syst. Tech. J.* **8** (1929), 613–31, and the subject was expanded by Neyman & Pearson in 1933. (I am indebted to Prof. Wilks for the first of these references.) As pointed out by Prof. Barnard at a recent British Mathematical Colloquium, the notion of errors of the first and second kinds ignores questions of 'robustness' of a significance test. One wants a test to be sensitive in detecting certain types of departure from the null hypothesis but insensitive to other types of departure, a compromise between robustness and sensitivity.

(x) *The point of a significance test.* What is the point of a significance test anyway? A large enough sample will usually lead to the rejection of almost any null hypothesis (cf. *Probability and the Weighing of Evidence*, p. 90). Why bother to carry out a statistical experiment to test a null hypothesis if it is known in advance that the hypothesis cannot be exactly true? The only answer is that we wish to test whether the hypothesis is in some sense approximately true, or whether it is rejectable on the sort of size of sample that we intend to take. These points are not usually made clear in text-books on statistics, and in any event they have never been formulated precisely.

(xi) *Is every significance test also an estimation problem?* This is another question on which there is controversy, but for our present purposes it is rather a side issue.

(xii) *On the use of random sampling numbers.* In avoiding the use of the Bayes–Laplace philosophy or of the modified Bayes–Laplace philosophy, orthodox statisticians attempt to make use of only two types of probability. These are (*a*) the tautological ones that occur in the definition of simple statistical hypotheses, and (*b*) probabilities obtained from random sampling numbers or roulette wheels or in some other way that does not lead in practice

to much dispute concerning the numerical values of the probabilities. (In effect, nearly everybody accepts the null hypothesis that the random sampling numbers are at least approximately equiprobably random. This is an example where the distinction, made, for example, by Fisher, between non-rejection and acceptance seems to disappear.) The usefulness of the method of randomization in the design of an experiment is indisputable; nevertheless, it becomes controversial if it is put forward as absolutely precise.

We consider the famous tea-testing experiment (see R. A. Fisher, *The Design of Experiments* (1949), p. 11). A lady claims to be able to tell by tasting whether the milk or tea is put first into a cup. An experiment is carried out consisting of twenty trials, made up of ten M's and ten T's, where M means 'milk in first' and T means 'tea in first'. By means of random sampling numbers it is arranged that all $20!/10!10!$ sequences are equally probable. By this design it is possible to make precise probability statements about how likely it is that the lady will make any given number of correct statements, assuming the null hypothesis; namely, that she is deluded.

Suppose we choose the sequence $MMMMMMMMMMTTTTTTTTTT$, and suppose that the lady gets all of her twenty statements correct. Do we really believe that her success should be measured by a tail-area probability of $1/2^{20} < 1/1{,}000{,}000$? For all we know the chance that she will guess that the above order has occurred, for the wrong reason, is far greater than this. (In fact, on the only occasion that I have seen this experiment performed, the statistician actually cheated and used the above sequence to save himself the trouble of randomizing.) A fluke of one in a million has occurred (if the statistician was honest), but we do not know how much of it is relevant to the main question of whether the lady can tell, and how much of it merely resides in the particular sequence of M's and T's that happened to be selected. We can, of course, use restricted randomization, so as to exclude the simplest sequences, but we can never entirely overcome the objection.

Thus the precision obtained by the method of randomization can be obtained only by ignoring information; namely, the information of what particular random numbers (or Latin square, etc.) occurred.

(xiii) *Does decision theory cover ordinary inference?* Just as there is fairly general agreement about the direct probabilities arising from random sampling numbers, there is also fairly general agreement within firms concerning certain utilities that occur in industrial processes. This is so when the utilities can be expressed in monetary terms and when the amounts of money are not large compared with the total capital of the firm. But in purely scientific matters there is much less agreement; in fact, the utilities as judged by a single individual will probably be bounded by upper and lower bounds that are very unequal. In other words the utilities are vague. For this reason the application of decision theory to scientific research is controversial (cf. *Probability and the Weighing of Evidence*, p. 40). But it does seem possible, after all, to apply decision theory quite sensibly to pure science, for example, by the use of the Type II minimax principle, which is an attempt to achieve precision when judgments are vague (*J. R. Statist. Soc.* B, **17** (1955), 195–6).

This note is based on a lecture given to the American Statistical Association and to the Society of Industrial Applied Mathematics, New York. The copyright is held by General Electric Company, who have kindly granted permission to publish. The present version gives effect to improvements suggested by Mr Wilfred Perks.