Statistical Inference

Dennis V Lindley - Somerset, England

1. Probability

My thesis is simply stated:

> The core concept, around which all statistics teaching should be based, is probability.

I do not mean by this that ideas like data analysis or graphical displays should be neglected. On the contrary. But I do mean that their effectiveness should be judged by how well they help in understanding the probability structure of the situation under study.

Nor do I mean that we all have to become mathematicians, discoursing learnedly on limit theorems and abstract spaces of probability measures. This is entirely unnecessary except for a few specialists. Probability, like geometry, has two aspects. First there are the rules of probability and the theorems that follow from them. They are studied by probabilists, playing a role similar to geometers with their study of spatial problems. Second, there is the measurement of probabilities and the calculations that flow from them. This activity is pursued by statisticians. It is interesting that geometers rarely measure anything. This activity is left to surveyors, who measure their angles and calculate their plans and maps. Statisticians are more like surveyors than geometers. Of course, surveyors use the Euclidean rules of geometry: so statisticians use the rules of probability. But their role is that of measurers and calculators.

2. Uncertainty

Why should probability occupy such a central position? The answer to this question is in two parts. The first rests on the appreciation that:

Statistics is the study of uncertainty.

An individual contemplating the world about him is surrounded by uncertainty concerning that world. All of the future, most of the past, and a lot of the present are uncertain. Some of this uncertainty can be eliminated. A physicist has been quoted as saying he did not need statistics; he did a bigger experiment instead. Fine, but this is not always possible, and even when it is, not always economically sensible.

The statistician's approach is to accept uncertainty as part of life and to develop ways of handling it so that the individual can rest comfortably with the lack of sure knowledge. The aim in teaching statistics is to provide the pupils with tools for handling uncertainty so that they can understand and respect it instead of pretending, as many do, that it does not exist.

3. The inevitability of probability

Accepting that statistics is the study of uncertainty, the second part of our support for probability rests on the view that:

Probability is the only sensible measure of uncertainty.

It is natural, confronted with a notion of uncertainty, to try to measure it, for measurement has been so successful in understanding many things, although it has had failures. Many suggestions have been made for measuring uncertainty: probability, odds, likelihood, fuzziness, support, chaos, etc. These measures differ in respect of their rules of combination. Thus probabilities add (under some circumstances) and multiply (under others), whereas fuzzy statements combine using rules based on maxima and minima. As with geometry, it is the rules that come first. With these settled, it is possible to measure and then to calculate according to the rules. So the displayed sentence above would be better altered to read:

The rules for combining uncertainty statements must be the rules of probability.

A key-word in this sentence is 'must'. It is not commonly recognised (for example, by fuzzy folk or chaos enthusiasts) that the rules of probability, commonly called axioms, can be derived from weaker and more transparent axioms. The first derivation was provided by Ramsey. Later proofs have been given by Savage, Jeffreys and others. The simplest, which uses nothing more than elementary Euclidean geometry, is given by de Finetti (1974). I have spoken of "the inevitability of probability". Very straightforward and simple ideas, coupled with strict mathematical proofs, demonstrate that the rules of combination cannot be left to whim or fancy but must be those of probability.

Hence statistics must be built around probability because statistics is the science of uncertainty and the rules for uncertainty are those of probability.

4. Conditioning

Before considering these rules, an important feature of uncertainty, and hence of probability, must be noted.

Probability is a function of two arguments.

The key-word here is 'two'. It is obvious to all that probability depends on the uncertain event, or quantity, being considered. But it is often not recognised that it also depends on what the individual making the probability judgement knows at the time. For an uncertain event A, we often talk of the probability of A and write p(A). But that uncertainty can change, not because A changes, but because the information changes. If your knowledge at the time you contemplate A is denoted by B, we should write $p(A|B)$ and speak of the probability of A, given B; a function of two arguments, A and B.

Here is an example. Consider a 50-year-old, British male and the uncertain event that he will die within a decade. You might arrive at a probability by consulting the actuarial tables. But suppose you learn that he has lung cancer. Your probability increases and the actuarial value is of little help except as a lower limit. If, instead, you learn that he belongs to a profession and that a grandfather and both parents are still alive, your probability will diminish. In all three cases, the uncertain event A remains 'death within the decade'; only what you are given, B, changes.

Sometimes the conditions B include supposition as well as fact. For example, statisticians often refer to the probability of data x given the value of a parameter θ, even though they do not know the value of θ. It turns out that it is unnecessary, in this context, to separate supposition from fact. In $p(A|B)$, B can include both.

The form $p(A|B)$ is sometimes called the conditional probability of A on conditions B, to distinguish it from probability, p(A). Since we do not admit the latter, the qualification on the former is not needed.

5. The rules of probability

The rules of probability can now be stated. There are three basic ones from which all others can be derived (with a qualification noted later).

Convexity	$p(A|B) \geq 0$ and $p(A|A) = 1$
Addition	$p(A \cup B|C) = p(A|C) + p(B|C) - p(AB|C)$
Multiplication	$p(AB|C) = p(B|C)p(A|BC)$

Here A, B and C are three events; $A \cup B$ denotes the union of A and B; AB, sometimes written $A \cap B$, denotes the intersection of A and B.

The convexity rule merely establishes the range (0,1) for probability with the upper end denoting certainty, for if you know A to be true, A is certain. The addition rule is the familiar one. It is usually stated for exclusive events; that is, AB is impossible, when

$$p(A \cup B|C) = p(A|C) + p(B|C).$$

The only rule with any subtlety to it is the multiplication rule. It is the only one of the three in which the conditioning event changes (from C to BC) and consequently is the only one that requires the two-argument notation. The study of probability that has become popular over the last fifty years has played down the multiplication rule by treating probability as a normed (convexity) measure (addition). To do this buries the vital idea, contained in the last rule, of how your views change on the receipt of additional information that B is true beyond your original knowledge of C. It is essentially Bayes rule, describing how we learn from experience.

As just emphasised, these rules can be proved. However, the addition rule is ordinarily employed in a form that cannot be proved. It is usually assumed to hold for an infinity, and not just a finite number, of events.

General addition For a sequence $A_1, A_2, \ldots$ of mutually exclusive events, $p(\cup_i A_i) = \Sigma_i p(A_i)$.

Without the generalisation, some 'obvious' results are not necessarily true. The general form need not be taught to any but mathematically sophisticated students. The three basic rules should be understood by everybody because they underpin any sensible appreciation of this uncertain world that we occupy.

6. Probability as belief

Rules alone are not enough. One must be able to interpret the probabilities, just as one has to think of Euclid's points and lines as marks on paper or something similar.

> Probability $p(A|B)$ is your measure of your belief in the truth of A when you know, or suppose, B to be true.

There are two key ideas here. The first is that probability is a property of an individual - here called 'you'. It is sometimes said to be *subjective* - a property of a subject - or *personal*. There is no reason why two people should not have different beliefs in A even when what they know is the same. This accords with practice. Even experimental scientists differ in their beliefs in a theory when the data are inadequate. Recent studies of the greenhouse effect provide an example. As suitable data accumulates, they come closer to agreement (using the multiplication rule) so that ultimately there is an appearance of objectivity that is held to be the hall-mark of scientific thought. De Finetti put it cleverly in his aphorism, "Probability does not exist", by which he meant it has no existence outside of an individual. It is neither a property of the world, nor of a person. It expresses a relationship between a person and the world.

The second key notion is that of belief. It is often held that probability is a frequency concept. This is incorrect and arises because of a confusion between the uncertainty itself and the data that might help in its measurement. That data is often a frequency. In the example of the 50-year-old, British male cited earlier, reference was made to actuarial, frequency data. Suppose 12% of such persons have been observed to die within the decade. Then this data may lead you to a probability of 0.12, representing your belief that John Smith will die. But there is no logic that impels you to do so. You may feel the observations are biased by many deaths being excluded and, as a result, choose 0.14 as your probability. The frequency and the probability are logically separate. You may up your belief to 0.80 on learning about the lung cancer despite the lack of frequency data. Only the belief is useful in connection with A, John Smith's death. The data may help in your measurement of that belief. It would not be wrong to have a belief of 0.20, instead of the frequency 0.12. Under mild conditions, as the data base increases, you will, again by the multiplication rule, tend to the frequency value of 0.12.

In this view, probability is very democratic and free, for you can believe what you like. But it is very demanding and restrictive in that your beliefs must fit according to the inviolate rules of probability.

7. Coherence

The fact that beliefs must satisfy probability rules is important and can be expressed in the aphorism:

Coherence is all.

A person is said to be coherent (or, more correctly, their beliefs are said to be coherent - the extension to action will be discussed later) if their uncertainties combine according to the rules of probability. Of a plant whose flowers can only be red, blue or white, a person whose probabilities for red, blue and coloured are respectively 0.3, 0.4 and 0.8 is incoherent because the last is not the sum of the other two, as the addition rule requires. That coherence is all, means that coherence is the only constraint. Any three values in which the last is the sum of the other two are permissible.

Coherence is an important tool in the measurement of beliefs. Here is a familiar, simple example. Most people will have probability 1/365 that a stranger will have their birthday on a specific date, year unspecified. Most people will judge probabilities for different strangers to be independent. Coherence requires that the probability that no two strangers in a room of 23 will share a birthday is about 1/2. The general idea is that some probabilities are easily measured, usually because there is, as here, frequency data to support them. From these values, using the rules, others, like that of the 23 people, may be found. A similar situation holds in surveying. One base line and many angles are easy to measure. From them, using the coherence of geometry, the distance from London to Edinburgh can be calculated.

8. Bayes theorem

Since uncertainty is omnipresent, and uncertainty must be measured by probability, it follows that:

Everyone should be taught the rules for probability.

I have no experience of teaching outside of universities and it would be presumptious of me to say how this can best be done. But there are some properties of probability that are simple, understandable and of practical value. Bayes theorem (or rule) is perhaps the most important. The multiplication rule says

$$p(AB|C) = p(B|C)p(A|BC).$$

Interchanging A and B has no effect on the left-hand side but the right becomes $p(A|C)p(B|AC)$, which must therefore be equal to the original right-hand side. Assuming $p(B|C) \neq 0$, this equation gives

$$p(A|BC) = p(A|C)p(B|AC)/p(B|C). \tag{1}$$

This is the theorem. Its importance lies in relating $p(A|C)$ to $p(A|BC)$, showing how the uncertainty of A is changed by the knowledge of B (in addition to C).

> Bayes theorem explains how we should learn from experience.

It also has importance in switching events. It relates $p(A|BC)$ on the left with $p(B|AC)$ on the right, interchanging A and B. People often have trouble distinguishing between these uncertainties, yet they are logically, and can be numerically, very different.

The theorem becomes easier to appreciate if, included with the result above, is that obtained by replacing A by its complement A^c. On dividing each side of (1) by the corresponding side of the new result, we obtain

$$\frac{p(A|BC)}{p(A^c|BC)} = \frac{p(B|AC)p(A|C)}{p(B|A^cC)p(A^c|C)},$$

eliminating $p(B|C)$. If C is omitted from the notation, because it plays a constant role in every conditioning event, and the odds notation o(.) is used, Bayes theorem says

$$o(A|B) = \frac{p(B|A)}{p(B|A^c)}\, o(A).$$

In words, the original odds on A are multiplied by the likelihood ratio $p(B|A)/p(B|A^c)$ to obtain the final odds given B (in addition to C). The introduction of logarithms makes the result additive.

Here is an example of an important, non-quantitative lesson that can be learned from the theorem.

> In judging the effect of data B on some theory A, account must be taken, not only of the probability of B were the theory true, but also were it false.

The likelihood ratio that multiplies the original odds to obtain the new odds is a ratio of the two probabilities. Thus in a court of law where A is the event of the defendant's guilt, and A^c innocence, the probability of evidence B both on the assumption of guilt, and of innocence, need to be compared. The result also shows how unsound popular, tail-area, significance tests are when, as almost always happens, only the significance level is quoted. This level is a probability on the assumption that A is true. What happens when it is false is ignored.

9. Independence

An important idea in connection with beliefs is that of independence. Two events, A and B, (or more correctly, the beliefs about two events) are independent, given C, if $p(A|BC) = p(A|C)$. In words, given C, the uncertainty of A is not altered on learning that B is true. Other ways of expressing the same idea are $p(AB|C) = p(A|C)p(B|C)$, which is symmetric in A and B and follows from the multiplication rule; or $p(B|AC) = p(B|A^cC)$, from Bayes theorem.

One often sees the statement that A and B are independent. This is unsatisfactory because the conditioning event is omitted and so fails to recognise that probability depends on two arguments. It comes as a surprise to many people to learn that A and B can be independent given CD, and given CD^c, but not independent given C alone.

10. Simpson's paradox

This last idea extends to include Simpson's paradox which is of great practical importance. Here is an example. T refers to patients given a treatment, T^c to those given a placebo (the controls).

Males	R	R^c	Rate	Females	R	R^c	Rate	Overall	R	R^c	Rate
T	18	12	60%	T	2	8	20%	T	20	20	50%
T^c	7	3	70%	T^c	9	21	30%	T^c	16	24	40%

R means recovery, R^c means no recovery. The entries are numbers of patients in the various classes. Three tables are given: for males only, for females only, and for all, obtained by adding the entries in the other tables. As judged by the recovery rates, the treatment is effective overall, but deleterious both for the men and for the women. Had you been given only the overall table that disregards sex, you might have concluded that the treatment was beneficial. In fact, it is not. On measuring probabilities by rates (equivalent to frequencies) the paradox translates easily into probability terms.

The paradox demonstrates the dangers of reaching conclusions based on inadequate data. (The inadequacy here is the failure to consider sex.) The difficulty is caused by the confounding of sex with the treatment. The recovery rate for females is much less than for males, yet males predominantly received the treatment, the females the placebo. Consequently, overall the treatment appeared to do well because it was mainly used on the men. The paradox can lead elegantly into a discussion of experimental design and the dangers of survey data. And all this with little mathematics beyond simple probability manipulations.

11. Decision-making

Statistics, both in teaching and in practice, is dominated by the ideas of inference; or what we have here termed beliefs. In the example just given, a statistician might infer, or believe, the treatment was harmful. But why have beliefs? Surely as a basis for action. Actions speak louder than words. Probability, as developed here, is admirably suited to the problem of choosing amongst actions; or decision-making as it is usually called, no distinction being made between the decision to act and the actions themselves. Indeed, both Ramsey and Savage derived probability directly from decision considerations. De Finetti's method can be regarded as using an action to choose a number to describe uncertainty.

Consider a decision d and an uncertain event A. For example, d might be the decision to go on a picnic and A the event of rain. The pair (d,A) is called a consequence; what will happen at a picnic spoilt by rain. The key idea is to associate with every consequence a number, called its utility, that measures the worth to you of the consequence. It is a simple matter to derive such a measure in terms of probability, so that utility obeys some of the rules of probability. Because of this observance, the rules can be used to show that the best decision is that which maximises your expected utility. In the simple case of one uncertain event, the expectation is

$$u(d,A)p(A|B) + u(d,A^c)p(A^c|B),$$

where B is the knowledge you possess at the time the decision about the picnic is to be made, and $u(d,A^c)$ is the utility in the event of no rain. Strictly both utilities should also refer to B since a change in B could affect your appreciation of the picnic. The expectation has to be compared with those of other possible actions, like a visit to the cinema. Details will be found in the text, Lindley (1985).

12. Comparisons

By introducing probability and utility in the way suggested, it should prove possible to explain the principles of sound (or coherent) beliefs and decision-making to everyone. As an example of the principles that emerge, we have:

Beliefs and actions are always comparative.

By this is meant that an action should not be selected solely on the basis of its own features, but in comparison with other possible actions. One does something, not because it is good, but because it is better than anything else that has been considered. We saw a similar feature in the legal scenario. Evidence must be judged by comparison of how probable it is on the assumptions of innocence and of guilt. There are no absolutes in this world, only contrasts. Is a probability more than another? Has this action higher utility than that?

13. Conclusions

Let me conclude by adding a second thesis to my initial one about probability:

The basic ideas of probability and utility should be taught to everyone.

The reason for this desire that all be exposed to these ideas is that every citizen is going to have to face an uncertain world; he or she is forced to make decisions. It is only recently that we have learned how to handle uncertainty, or how to make individual decisions in a coherent way. This knowledge must be shared for mankind effectively to handle the difficult problems ahead. These ideas do not tell us how conflict can be resolved but they do at least enable people who are collectively in broad agreement to act more sensibly than they would without probability concepts. We have been too narrow in our perception of what statistics can do. The subject is not just about numbers, nor even conclusions based on numbers, but rather it is a way of organising one's thinking and acting in an uncertain world.

References

De Finetti, B (1974) *Theory of Probability* (Vol 1). John Wiley & Sons, London.
Lindley, D V (1985) *Making Decisions* (2nd ed). John Wiley & Sons, London.