Archives of Disease in Childhood 1994; **71**: 90–94

STATISTICS FROM THE INSIDE

13. Probability and decisions

M J R Healy

In this series I have routinely used the language of probability in describing the results of statistical analyses. I have talked of the probability of errors of the first and second kind when designing a study; both significance tests and confidence intervals have been defined within a probability framework. It will be worth spending a little time considering what probability is and how probability statements should be interpreted.

To get one problem of terminology out of the way, strictly I should speak of the probability attached to statements or propositions, such as 'the patient will recover if given this treatment'. It is convenient though to speak also of the probability of the event which the statement is about, here the patient's recovery. If A denotes some event I shall write pr(A) to stand for the number which is the event's probability.

The meaning of probability statements

There are two schools of thought concerning the meaning of a probability statement which have coexisted throughout the 500 year history of the concept. For one school, the subjectivists, a probability is a numerical assessment of somebody's *degree of belief* in a certain statement. The two extremes of belief are impossibility – the statement cannot possibly be true – and inevitability – the statement is quite certainly true – and it is convenient to give these the values 0 and 1, so that the probability of occurrence of an event about which I am uncertain will be a number between 0 and 1. It is important to note the appearance of the personal pronoun in the preceding sentence. In this interpretation, a probability statement measures the degree to which some proposition is believed by someone; probability is essentially something subjective. It is entirely legitimate, and may be very reasonable, for your probability of a certain event to be different from mine. This is partly because we are both entitled to our own opinions, but also because of the important consideration that we may possess different amounts of information concerning the event in question.

Suppose for example that A is a diagnostic statement: 'this patient is suffering from condition X'. An experienced clinician in an outpatient clinic will be able to attach a probability to this statement before the patient comes through the door – she will know that in her clinic condition X is rarely seen and that it is quite unlikely to turn up in the next patient on the list. But after taking a history and making an examination, the clinician's degree of belief that the patient is suffering from condition X will change; in favourable circumstances, the diagnosis of condition X will become almost certain or almost ruled out, so that the probability attached to it will approach 1 or 0.

Suppose that A is a statement and that H stands for a set of other statements (derived from the history, test results, etc) which are to a greater or lesser degree relevant to A. We write pr(A|H) for the probability of A given the knowledge that the statements H are true. This is called a *conditional probability*, the probability of A conditional upon H. The point I have made above about the changes in probability with the accrual of evidence amounts to the important doctrine that *all probabilities are conditional*, they all depend upon what evidence is available. If for simplicity's sake I write pr(A) rather than explicitly pr(A|H), this must mean that all those addressing the problem have H in common, are possessed, that is, of the same items of evidence. We shall see that under these conditions, everybody should have almost the same degree of belief, so that probability is less subjective than it may seem at first. Note for future reference the somewhat surprising appearance of the word 'should' in the preceding sentence.

The rival *frequentist* interpretation of probability statements relies upon the notion of relative frequency. This theory grew up in the context of gambling where it arises very naturally. Suppose the event we are talking about occurs as one in a set of identical independent trials, each trial having an outcome which can be classified as a success or a failure. Obvious examples which proliferate in the textbooks are tosses of a coin, throws of a dice, or spins of a roulette wheel. With a dice for example, suppose we classify a 6 as a success and other outcomes as failures. The physical symmetry of the dice suggests that each of the six faces will turn up equally frequently in the long run. It follows that the event will count as a success in one sixth of all trials, and we therefore assert that the probability of the event is 1/6.

This frequency theory of probability seems at first sight a good deal more solid than the wishy-washy subjectivist theory, more scientific perhaps. A closer look is rather

23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD

Correspondence to:
Professor Healy.
No reprints available.

disillusioning. There are two key phrases in the above description, 'independent' and 'in the long run'; no really satisfactory explication of the latter has ever been produced, while the former seems only to be definable in terms of probability, the concept that is supposed to rest upon it. As well as this, the theory is all very well for gamblers sitting interminably at the casino table, but in a clinical environment the notion of a long series of identical trials is a good deal less acceptable. For the recipients of genetic counselling, conscious of their unique situation, long term frequencies may have no appeal at all.

The frequentist theory has been dominant in statistical circles for the past 70 years or so and underlies all the standard statistical methodology that is taught in medical schools and which I have been expounding in these notes. A statement, for example, that a population mean lies within a certain 95% confidence interval is expected to be true with a long run frequency of 0·95. In recent times, many statisticians have become more and more dissatisfied with it and have favoured the subjectivist theory in its place. It has to be said that the two approaches are to some extent complementary. The frequentist must have subjective beliefs if the theory is to have any applicability – unless we believe that frequently occurring events are likely to occur on a particular occasion, the theory is sterile. On the other hand, the subjectivist needs relative frequencies for the purpose of calibration. To say that my degree of belief in some diagnostic statement is 5/6 is not in itself very meaningful, but if I say that it is equal to my degree of belief that the next throw of a dice will not produce a 6 the interpretation is a good deal more immediate.

Combined probabilities

The key question of probability theory is how to combine probabilities. Given the probabilities of two events A and B, what is the probability that one or other of them occurs, or that both of them occur? The two theories agree on the answers to this question and the frequency theory makes it quite easy to establish the necessary formulas, but the approach via the subjective theory is more interesting. It is rather odd, though, that such an approach is possible at all. If probability is subjective, how can it make sense to ask what the answer to a probability question *should* be? Is it not open to you to entertain any opinion you like as to the probability of (say) A or B occurring? Certainly it is; but one particular opinion turns out to be more sensible than alternatives.

It seems sensible, for instance, that one's probability judgments should be *coherent*. This technical term means that if I judge that event A is more probable than event B and also that event B is more probable than event C, then I also judge that event A is more probable than event C. The reasonableness of this kind of guideline is fairly obvious if for 'more probable' you substitute 'heavier' or 'louder'. But a more down-to-earth argument in favour of coherency can be adduced. Suppose your* probability judgments are non-coherent – you reckon A more probable than B, B more probable than C, and C more probable than A – and that you are prepared to back your probability judgments financially, to put your money where your mouth is. Then you will be prepared to bet with me that A will occur as against B, that B will occur as against C, and that C will occur as against A. For definiteness, let all the bets be at £2 to £1. Now suppose that event A actually occurs. You win your first bet and receive £1; the second bet is void; and you lose £2 on the third bet, finishing up £1 out of pocket. But clearly the outcome is the same no matter which event occurs. A person who is confident in non-coherent probability judgments is a potential financial disaster waiting to happen. This makes it the more remarkable that when a person's subjective probabilities are elicited (by no means an easy task), they usually turn out to be non-coherent.

It is essentially this result that forms the basis for the breathtaking claim made by exponents of the subjectivist theory that probability along with its laws based solely upon the notion of coherence forms the only correct way to reason about uncertainty. Other alternatives, such as fuzzy logic or the varieties of so-called intelligent systems, must be either equivalent or demonstrably inferior.

The laws of probability

What then are the laws for combining probabilities that the notion of coherence leads to? There are two basic ones, both very simple. Given two events A and B, what is the probability that one or the other or both of them occur? If we simply add the individual probabilities pr(A) and pr(B), it turns out that we are counting the double occurrence A−and−B twice, so that the correct answer can be written as

$$\text{pr}(\text{A}-\text{or}-\text{B}) = \text{pr}(\text{A}) + \text{pr}(\text{B}) - \text{pr}(\text{A}-\text{and}-\text{B})$$

In a common special case, the events A and B are mutually exclusive so that they cannot both occur. Then the double event A−and−B is impossible, the last term in the above formula is equal to zero, and the joint probability is just the sum of the individual probabilities. This is the so-called *addition law*.

If A and B are not mutually exclusive, what is the probability that they both occur? The answer involves the use of conditional probabilities that I introduced earlier. It can be found by considering one event to occur before the other. We get

$$\begin{aligned}\text{pr}(\text{A}-\text{and}-\text{B}) &= \text{pr}(\text{A}) \times \text{pr}(\text{B}|\text{A}) \\ &= \text{pr}(\text{A}) \times \text{pr}(\text{A}|\text{B})\end{aligned}$$

where p(A|B), for example, means the probability of A when I know that B has occurred.

*The characters 'I' and 'you' occur frequently in the subjectivist probability literature, where they play the part of principal and stooge respectively.

Once again there is an important special case. Suppose pr(A|B)=pr(A), that is to say that the probability of A is the same whether B occurs or not. We then say that A and B are *statistically independent*, and for statistically independent events we have

$$\text{pr(A-and-B)}=\text{pr(A)}\times\text{pr(B)}$$

the *multiplication law*.

Bayes' theorem

The double equality shown above has an important consequence. Let me denote the presence of a disease by D and the occurrence of a relevant positive test result by T. What is the probability of a patient with the disease turning up and giving a positive test? We have

$$\text{pr(D-and-T)}=\text{pr(D)}\times\text{pr(T|D)}=\text{pr(T)}\times\text{pr(D|T)}$$

This can be rewritten in two slightly different ways. First

$$\text{pr(D|T)}=\text{pr(T|D)}\times\{\text{pr(D)/pr(T)}\}$$

This shows how to flip the two components of a conditional probability. The quantity pr(T|D) is what is usually called the *sensitivity* of the test, the probability that a patient with the disease will produce a positive test result. But this is not what the clinician facing a test result wants to know. She is interested in pr(D|T), the so-called *positive predictive value*, which is the probability that a patient with a positive test result actually has the disease. The formula shows that the conversion from one to the other involves pr(D), the pretest probability of the disease condition, in other words the *prevalence*. I intend to discuss this and related questions further in a later article in this series.

The other useful form of the double equality is

$$\text{pr(D|T)}=\text{pr(D)}\times\{\text{pr(T|D)/pr(T)}\}$$

This is the famous *Bayes' theorem*, which promises no less than a recipe for how the sensible person should learn from experience. I start with pr(D), my present degree of belief that my patient is suffering from the disease condition. I then acquire a new piece of evidence T in the form of the test result. The formula then shows me how, if I am to remain coherent in my beliefs, I must change my degree of belief to pr(D|T). The terms pr(D), pr(D|T) are called the prior and posterior probabilities (prior and posterior, that is, to the acquisition of the piece of evidence T) and the multiplying factor is called the likelihood.

Bayes' theorem can be expressed in another form which is of interest. Consider not-D, the proposition that the patient did not have the disease. Note that as D and not-D are mutually exclusive and one or the other must occur, pr(D)+pr(not-D)=1, so pr(not-D)=1-pr(D). From Bayes' theorem we have

$$\text{pr(not-D|T)}=\text{pr(not-D)}\times\{\text{pr(T|not-D)/pr(T)}\}$$

and dividing this into the previous equation we get

$$\frac{\text{pr (D|T)}}{\text{pr(not-D|T)}}=\frac{\text{pr(D)}}{\text{pr(not-D)}}\times\frac{\text{pr(T|D)}}{\text{pr(T|not-D)}}$$

Now if p is some probability or other, the quantity p/(1-p) is the corresponding *odds* – instead of talking about a probability of 1/6 we can talk about odds of 1 to 5 or 5 to 1 against. Then pr(D)/pr(not-D) is the prior odds for the event D and pr(D|T)/pr(not-D|T) is the posterior odds given the event T, the positive test result. To go from one to the other we have to multiply by the ratio pr(T|D)/pr(T|not-D). This might be called the selectivity of the test but it is usually known as the *likelihood ratio*. If we take logs of each term in the above formula the multiplication is replaced by an addition and we see that performing the test has increased the prior log odds by adding a certain amount. An alternative test would add a different amount depending upon the relevant probabilities. The log of the likelihood ratio is thus an additive quantity which measures the *weight of the evidence* that the test provides. Alongside financial costs and other considerations, it can be used to decide which of two alternative tests is to be preferred.

Suppose that you and I hold different opinions concerning a proposed new treatment for some disease condition. These different opinions will lead us to have different probabilities for the effectiveness of the treatment in a forthcoming case. In order to use the theory, we both need to express these probabilities in numerical form. I have little information about the new treatment and I believe it is as likely to succeed as to fail; I can express my state of belief by equating it with having observed (say) five successes and five failures in previous experience. You are distinctly sceptical and firmer in your opinions; you equate your state of mind to having observed (say) 10 successes and 20 failures. Now suppose that we both acquire knowledge of the results of a clinical trial in which 120 successes and 40 failures of the treatment were observed in cases comparable to the one under discussion. Then to remain coherent in our beliefs we must both of us update our probabilities using Bayes' theorem. This is very simple; we should now assess our probabilities by adding the actual cases from the trial to the hypothetical ones that we envisaged *a priori*. Thus I should behave as though I had observed 125 successes and 45 failures, while you should behave as though you had observed 130 successes and 60 failures. While our opinions still differ, we both agree that the treatment is on the whole effective and the extent of our disagreement had been considerably reduced. It is reassuring that, even assuming a subjective nature for probabilities, coherence leads to evidence from actual data modifying and eventually overcoming initial differences in degrees of belief. When a frequency probability is meaningful, coherent observers will adjust their subjective degrees of belief until they come to agree with it.

There is an exception to the above result, though. Suppose I am absolutely certain that the new treatment is ineffective and that my prior degree of belief in its effectiveness is represented by a probability of zero. Then an inspection of the Bayes' theorem formula will show that this zero probability will not be modified by any amount of actual data. In this context, 'certainty' means what it says; my mind is made up and is not to be confused by facts. Anyone is at liberty to possess *a priori* probabilities of 0 or 1, but they should realise the consequences of such degrees of belief. As Cromwell said to the elders of the church in Scotland, 'I beseech you, in the bowels of Christ, think it possible that you may be mistaken'. One way to do this is to think in terms of odds rather than probabilities; there is no harm in prior odds of 100 or 1000 to 1, as even these can be modified by a sufficient quantity of data.

The explicit introduction of prior probabilities has implications for the practice of medical statistics which are only just being explored. The standard method for determining sample size in a clinical trial may be regarded as a confidence trick; the user is told by the statisticians to specify the size and power of the appropriate significance test, but little or no guidance is given as to how these quantities should be determined. Suppose the spectrum of prior beliefs concerning the new treatment were to be investigated in a sample of interested professionals, then the amount of data that would be needed to bring these *a priori* beliefs into agreement could be calculated and would represent the appropriate size for the trial. This approach has been explored and is turning out to be a practical one. Another problem which has been difficult for the classical approach to handle is that of stopping a clinical trial when the initial results suggest that the new treatment is not effective (ordinary sequential trial theory considers the opposite problem of stopping a trial because the new treatment has proved itself to be effective). At a given point in the conduct of a trial, the results to date can play the role of *a priori* information. Based upon this, it is possible to evaluate the probability of achieving a significant outcome when the scheduled number of patients have been studied. If this probability is low, a strong case can be made for terminating the trial forthwith. More detail on these and other related topics can be found in Spiegelhalter *et al.*[1–3]

Table 1 Uncertain statements

	p%	*t*	*Penalty*
1. Edinburgh lies south of Copenhagen			
2. Iron is denser than copper			
3. St Sebastien is the patron saint of athletes			
4. Chaucer was born before Petrarch			
5. Trollope wrote more than 40 novels			
6. Raphael died before Michelangelo			
7. The frequency of middle C is approximately 520 Hz			
8. Mozart wrote more symphonies than string quartets			
9. St Paul's Cathedral, London is taller than St Peter's, Rome			
10. A nanosecond is approximately 1 light foot			
	Total penalty score		

Significance levels and degrees of belief

It would be natural to suppose that one's degree of belief in a hypothesis could be related to some form of significance probability – the more significant the experimental results, the less the degree of belief one should have in the null hypothesis. Remarkably, this is by no means true.[4] Suppose you have observed some data and are about to carry out a significance test. You will proceed to evaluate the probability of obtaining data as extreme as those observed on the assumption that the null hypothesis is true, and it follows that you must have some degree of prior belief that this assumption is correct – that the true difference between the mean responses to the treatment and the control, for example, is *exactly* zero. Suppose now that you have a lot of data (the sample means are very precise) and that the actual significance level is beyond the magic 5% level but not by very much. Then it is not difficult to show that the sensible person, updating their prior beliefs using Bayes' theorem, should actually be *more* confident that the null hypothesis is true, not less. Speaking extremely informally, if the null hypothesis is actually false and the data are as precise as this, the significance level has no business to be so unimpressive. The practical implication seems to be that we should vary our significance levels according to the sample sizes – a result just significant at the 5% level may be quite convincing with small samples, but with bigger ones a more stringent level is appropriate.

Probability assessment

Assessing degrees of belief in numerical terms is a skill that is by no means innate but has to be acquired. As a not quite trivial pursuit, you may like to try the following quiz. Table 1 contains 10 factual statements, some true, some false. You are asked to write against each statement the probability p that you personally attribute to it, in percentage terms. If you are absolutely sure it is true, write 99 (rather than 100, because of Cromwell's principle); if you are absolutely sure it is false, write 1. If you have no idea whether it is true or false, write 50; if you rather think it is false but are not too sure, write something between (say) 10 and 40. When you have finished, consult the key at the end of this article, and write the target score t against each of the statements, 100 for the true ones and 0 for the false ones. Then calculate the *penalty score* for each statement as $(p-t)^2$ and form the total. How did you do? If you were sure of all the answers *and were correct each time*, the total penalty would be 10; if you were sure and got each one wrong, the penalty would be 98 010. If you got bored and entered 50% for all the probabilities, you would get 25 000; if you put 1% or 99% at random without looking at the statements, you would get 49 010 on average (so that confessing your ignorance is better than guessing). Repeated exposure to quizzes of this type can lead to your probability assessments becoming *well calibrated*, in the sense that events to which you attribute a probability of p% tend to occur on p% of all occasions (notice the convergence

Table 2 Utilities

		Decision	
		Do not treat	*Treat*
True state of affairs	Disease absent	1·0	0·0
	Disease present	0·2	0·8

between the subjectivist and frequentist approached at this point).

Statistical decision theory

What use are probabilities, be they subjective or frequency based? One answer is that they can form the basis for practical decisions. As a non-clinician I hesitate to trespass, but it is obvious even to the outsider that clinical practice consists in repeated sequences of therapeutic and other decisions and that these decisions have to be made in the presence of uncertainty – even the most naive patients do not expect the doctor always to be certain exactly what is wrong with them and exactly what to do about it to make them better. If statisticians assert that probability is the optimal way of handling uncertainty, then they should come clean on just how this might be done.

Let us make the simplest possible model for a single decision (decisions normally come in chains, with each one depending upon the outcome of the previous one, but this would lead us too far astray for a short article). Suppose then that a patient may or may not have a certain fairly mild disease condition and that you have to decide whether or not to initiate a particular treatment which you know to be effective in this condition. In order to make the decision it is necessary to be quantitative about the various possible outcomes. Treating the patient when she does not have the disease may be regarded as the least desirable outcome and given a score of 0; not treating the patient and discovering that she does not have the disease the most desirable and given a score of 1. The other two outcomes must be given intermediate scores which will depend upon the detailed circumstances. A possible set of scores (the technical name is *utilities*) is shown in table 2. I want to emphasise that these are not numbers pulled out of the air; utilities in the statistician's sense are operationally definable quantities, though space forbids me to go into details. In particular, they should be coherent – if you prefer A to B and B to C, then you should prefer A to C.

Table 3 Utilities

		Decision	
		Do not treat	*Treat*
True state of affairs	Disease absent	1·0	0·0
	Disease present	0·3	0·6

The problem is clear. If the disease is absent, you prefer not to treat, if the disease is present, the other decision should be made. But you are uncertain as to which state of affairs obtains. What then is your probability for the presence of disease? Suppose that in the light of the information available you assess this as 0·7 – you suspect that the disease is present but you are by no means sure. Then you can work out the average or *expected utility* for each of the possible decisions. For 'do not treat' this is $0{\cdot}7\times0{\cdot}2+(1-0{\cdot}7)\times1{\cdot}0=0{\cdot}44$; for 'treat' it is $0{\cdot}7\times0{\cdot}8+(1-0{\cdot}7)\times0{\cdot}0=0{\cdot}56$. In spite of the uncertainty, with these preferences it pays to treat the patient.

But suppose for a different set of circumstances the preferences are expressed by the utilities in table 3. The disease is still mild but the treatment is unpleasant and of doubtful efficacy, so that the utilities in the second row of the table become 0·3 and 0·6. The average utilities are now 0·51 and 0·42. If you are really that uncertain that the disease is present, you should withhold treatment.

Undoubtedly what I have set forth is scarcely so much as a caricature of clinical reality. It is also a seriously incomplete account of a quite complex theory. Nevertheless it has some revealing features. Above all, it underlines the fact that arriving at decisions involves evaluating the relative desirability of possible outcomes. It follows from this that the 'optimal' decision may be different according to who does the evaluation. (In modern medicine there are at least three potential evaluators: the doctor, the patient, and the manager.) Moreover it brings out the role of uncertainty in determining a decision and the desirability of expressing this uncertainty in quantitative terms.

Further reading

Much of this article has been based upon a fascinating book by D V Lindley (*Making Decisions*, 2nd edition, Wiley 1985) which is strongly recommended to those wishing to investigate further. For a fuller account of statistical decision theory in a medical context, see the article by D J Spiegelhalter and A F M Smith in *Perspectives in Medical Statistics* (edited by J F Bithell and R Coppi, Academic Press 1981). For examples in medical practice see *Analysing How We Make Clinical Decisions* (edited by H Llewelyn and A Hopkins, Royal College of Physicians 1993) and articles in the journal *Medical Decision Making*.

[Quiz solutions: FFTFTTFTFT]

1 Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med* 1986; **5:** 1–13.
2 Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986; **5:** 421–33.
3 Spiegelhalter DJ, Freedman LS, Parmer MKB. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A* 1994; **157:** (in press).
4 Lindley DV. A statistical paradox. *Biometrika* 1959; **44:** 187–92.