# 10 Inferential decision

## DECISIVE AMOUNT OF EVIDENCE

I have used quotation marks to set off the word 'significance' and the phrase 'statistically significant' to signal that these terms (referring to an amount of evidence obtained in a test that we will accept as decisive) are easily misinterpreted. The confusion, I believe, stems from our desperate longing for anchors of certainty in a sea of doubt. We hope that if we go to the considerable trouble of conducting a comparative trial and observe the rules of evidence scrupulously, we will arrive at a point where we can stop and come to an unequivocal decision. We dream of moving from doubt to certitude about the questioned events, an action that is epitomized by the declaration, 'We have achieved a significant result.'

### Random error and systematic bias

Unfortunately, the telegraphic message conceals more than it reveals about the nature of verdicts made after clinical trials. For example, it may hide the fact that a 'significance' test is not a magical operation to determine whether or not there was some hidden bias in the investigation.

When a statistical test is applied like a recipe in a cook book, without first examining the experimental design used to reduce the influence of extraneous factors, the computation is a futile exercise (p 27). The mechanical practice reflects a commonly held opinion that effects revealed in experimental data are due only to factors under test plus chance (strictly random error). But a decision statement expressed in terms of probability has nothing useful to say about the role of chance unless we can be assured that random effects have not been overwhelmed by extraneous systematic influences.

*Terminology* If it were possible to outlaw the word 'significant' when describing statistical inferences, the ban would go a long way in improving the clarity of our assertions about the state of evidence in medicine. (Mainland once proposed that the ambiguous and grandiloquent term 'significance

test' be replaced by the more specific and explanatory title of 'random frequency test'.) But the unfortunate word is embedded in thought and in writing; little can be done except to sound warnings about the referents of the value-laden expression. Not the least of the difficulties is the confusion between 'statistical significance' and 'practical significance'; the two are not always congruent.

**The 0.05 threshold of 'significance'**

An illusory concept is that of a critical threshold of rarity (a 1-in-20 risk of Type I error stated as 'significant at the $\alpha = 0.05$ level'), as opposed to the notion of a gradient of improbability with no discernible 'break' in increasing odds against the likelihood that a specified event has occurred wholly by chance.

The use of verdict terms in statistical analyses of biologic problems was traced by Mainland to a paper written by Karl Pearson in 1896 in which differences in measurements of the human body were stated to be 'significant' or 'not significant' by reference to a quantity called the probable error. A 'significant' difference came to be defined as one that occurred rarely when random samples were taken from the same population, and 'rare' was defined later as outside the range of two standard deviations. This range excludes slightly less than 5 per cent of the total number of sample observations (about $2\frac{1}{2}$ per cent in each outlying 'tail' of the distribution, p 32).

R.A. Fisher adopted the custom and wrote, 'We shall not often be astray if we draw a conventional line at 0.05 and consider that ... [lower] values ... indicate a real discrepancy ... It is convenient to take this [0.05] point as a limit in judging whether a deviation is to be considered significant or not.'

Thus, the choice of the threshold level for Type I error owes more to historical custom than to some inflexible characteristic of data; the cut-off standard, it has been noted, it not a law of Nature. The choice of how often we are willing to 'be astray' is ours, and this determination cannot be made rationally without examining the possible consequences of our decision.

## SELECTION OF A 'SIGNIFICANCE' TEST

Before discussing the selection of standards of rarity, I should point out that the choice of the probability-measuring instrument—the 'significance' test—requires some thought. And these deliberations must take place *before* the collection of information begins. The assumptions concerning sampling distribution and sample size that underlie the commonly used tests are described in textbooks of statistics; and these pre-conditions guide the choice of a computational procedure for making statistical inferences.



**Badgering the data**

Nothing prevents us from performing arithmetic operations on any and all numbers obtained in a series of observations or a formal comparative trial. After the collection of data, we are not physically restrained from shopping for the 'significance' test that turns out the most impressive values when our numbers are plugged into the innocent formulae. But we cannot, in good conscience, lean on the laws of probability for support in the interpretation of such illegally derived results. A wag observed that almost any set of data, if sufficiently badgered, can be exhausted into submission.

**Parametric and nonparametric tests**

'Parametric significance test' is the term applied to a method that assumes some *particular* distribution of variables. (The word parameter refers to a specific value of a characteristic in a parent population.) The assumption is made that the frequency distribution of variables (height, for instance) in the parent population from which our patients are drawn has a bell-shaped symmetry (confusingly termed a 'normal' distribution). Although parametric tests are accurate in many situations in which characteristics of interest are not symmetrically distributed, the tests are weakened when the population distributions are markedly distorted and when very small sample sizes are analyzed.

Since these limiting conditions are commonly encountered in medical problems, nonparametric tests (which are developed without reference to the distributions of variables) have a distinct advantage. The distributions of age at death and the ordinally ranked grades of RLF are examples of the kinds of asymmetries which turn up frequently in event measurements. The most widely used nonparametric technique is the $\chi^2$ test for categorical data, that is, 'success', 'failure'; another is the Wilcoxon test for ranked ordinal data; these and others are described in statistical texts.

## CHOOSING A LEVEL OF 'SIGNIFICANCE'

The choice of a rarity level that will be called 'significant' in a clinical trial involves a thoroughgoing analysis of the actions contemplated by decision makers. In an applied field like medicine, we are obliged to consider the price we are willing to pay if our future actions turn out to be wrong. The gambling analogy is inescapable: prudent bettors examine their bankrolls and decide how much they are prepared to risk before setting the terms of a bet.

**Weighing consequences**

If the consequence of being wrong about a treatment involves a relatively small loss to future patients and to the community (such as frequent failure to shorten the course of a benign disease that usually subsides spontaneously, minor risk of treatment-induced complications, and small expenditure of personal and community resources), the risk level for Type I error may be set at, say, $P \leqslant 0.10$ (the capital letter $P$, for probability level, is the conventional label for the value). We can afford to take the risk of being wrong 10 per cent of the time when we declare that treatments of this kind are effective. As the potential cost of incorrect decisions increases, however, we wish to lower the risk of Type I error accordingly. The customary 5 per cent 'significance' level is too lenient for many situations in medicine.

*A concentration ceiling for oxygen treatment* The single hospital experiment concerning oxygen treatment, that I cited earlier (p 122), provides an instructive lesson of the need to weigh consequences even when a relatively conservative risk level for Type I error is chosen. At the completion of that trial it was observed that the rate of irreversible RLF was reduced from 22 per cent under 'high' oxygen treatment to zero among infants assigned to 'low' oxygen management. The difference in outcomes was declared 'significant at about the 2 per cent level' and it was concluded that RLF-blindness could be *entirely* eliminated by a regimen of strict regulation of oxygen treatment.

The investigators advised that supplemental oxygen should be given only when there were unequivocal signs of need and that a gas concentration of 40 per cent (the ceiling used in the 'low' oxygen group) should not be exceeded. The 'under 40 per cent' advice was widely publicized, and there was general acceptance of the view that eye damage would not occur if the concentration of oxygen in incubators was carefully monitored and kept below the 'danger' level. This erroneous notion persisted for years despite a growing number of observations that RLF-blindness was not eradicated by these stringent restrictions of oxygen concentration and that early mortality was increased.

It is unlikely that a larger trial would have changed the verdict that 'low' oxygen reduces the risk of blindness (differences as large as the one observed would be expected to occur by chance only once in 50 trials involving about 40 babies in each group). However, the limits of what we may safely infer from zero occurrence of RLF in a sample of 40 infants might have been appreciated if the report had included the statement 'We are 95 per cent confident that the true risk of blindness after 'low' oxygen treatment lies between zero and 9 per cent.'

Estimates of this kind are called 'confidence intervals'; they indicate the range from the smallest to the largest effect of a treatment (or difference between treatments) with which the trial results are consistent. A 95 per cent confidence interval, for example, is the 2 standard deviation range surrounding the observed value; it expresses our conviction that the interval covers the (unknown) true value in 95 per cent of all possible samples (p 32). The additional information supplements the results of 'significance' testing and should be provided in complete reporting of trial results.



**Trial size and claims of 'significance'**

A report by Peto and co-workers noted that the size of a trial should be considered in assessing claims of 'statistical significance.' It was observed that for every trial that compares two treatments for cancer that are substantially different in effectiveness, there are probably five to ten 'negative' trials in progress (the compared therapies are essentially equal).

**Numbers of trials that will be 'statistically significant'**

| *Planned trial size* | *Expected outcomes. Proportions of 'events' in two treatments* | *No. of trials*[a] | *Decision*[b] | |
|---|---|---|---|---|
| | | | *'Non-significant'* | *'Significant'* |
| About 250 'events' (enrollment of some hundreds of patients) | 50 × v. 50% (no real difference) | 100 | 95 (right) | At least 5[c] (misleading) |
| | 50% v. 33% (treatments really differ) | 20 | 1 (misleading) | 19[d] (right) |
| About 25 'events' (enrollment of some dozens of patients) | 50% v. 50% (no real difference) | 1000 | 950 (right) | At least 50[c] (misleading) |
| | 50% v. 33% (treatments really differ) | 200 | 150 (misleading) | 50[d] (right) |

[a] Numbers of such trials in progress throughout the world postulated by Peto's group.
[b] Given the postulated numbers, approximately how many will be declared 'nonsignificant', and how many 'significant' at $P<0.05$?
[c] Even the most rigorously designed, executed and analyzed studies have a 1-in-20 chance of a false positive result with $\alpha$ set at $P<0.05$. The less rigorous the study, the greater the chance of a false positive; in practice 10 per cent or more of all studies yield such false positives (at $P<0.05$).
[d] A reduction from 50 per cent to 33 per cent mortality has a very good chance of being detected (19 out of 20) in a trial in which hundreds are randomized and about 250 of them die, but the chances are only 1 out of 4 that a difference of this size will be recognized in a small trial in which dozens are randomized and about 25 succumb.
(Taken from Peto and co-workers)

Given this situation, a few reasonable numbers were postulated and three conclusions were suggested from the projections. First, a large proportion of reports of 'statistically significant' treatment differences in small trials are misleading: differences of the size predicted do not exist. Second, if a

small trial compares a new treatment that is so effective that it prevents one third of the deaths that occur under a control regimen, the study will probably fail to reach 'statistical significance'. Finally, a serious bias arises because most the interesting therapeutic questions are being studied simultaneously by many investigators. Any trials, large or small, in which patients given the new therapy fare significantly better will be published and publicized. The studies that find no difference will rarely receive wide attention, especially if the numbers of enrolled patients are small.

The inevitable bias, the report concluded, can be circumvented to some extent by restricting attention to medical trials so large that they would be published whether or not an important outcome difference was observed.

**Publication decisions**

... There's this desert prison, see, with an old prisoner, resigned to his life, and a young one just arrived. The young one talks constantly of escape, and, after a few months, he makes a break. He's gone a week, and then he's brought back by the guards. He's half dead, crazy with hunger and thirst. He describes how awful it was to the old prisoner. The endless stretches of sand, no oasis, no signs of life anywhere. The old prisoner listens for a while, then says, 'Yep. I know. I tried to escape myself, twenty years ago.' The young prisoner says, 'You did? Why didn't you tell me, all these months I was planning my escape? Why didn't you let me know it was impossible?' And the old prisoner shrugs, and says, 'So who publishes negative results?'

Jeffrey Hudson

## Publishing only 'significant' results

Theodore D. Sterling of the University of Cincinnati examined some of the issues concerning the publication of experimental results. He reviewed research reports in psychology research journals and found that of 294 articles using statistical tests, only 8 did not attain the 5 per cent 'significance' level. This evidence suggested that a fixed level of 'significance' was adopted as a criterion for submitting or accepting reports for publication. The practice, he noted, leads to unanticipated consequences.

When research that yields 'non significant' results is not published, the study may be repeated many times by investigators who are unaware of the previous efforts. Eventually, by chance, a 'statistically significant' result occurs—a Type I error—and this replication is published. Readers approach each report in a prestigious journal with the expectation that the results are 'statistically significant', but they should consider the possibility that a selection may have taken place. Among a set of similar experiments, the one that yielded large differences by chance had an artifically enhanced opportunity to come to their attention in print. Before readers can make an intelligent decision about what is published, they must have some information concerning the distribution of outcomes of similar experiments or at least some assurance that a similar test has never been performed. Since the needed information is unobtainable, the subscribers are in a quandary.



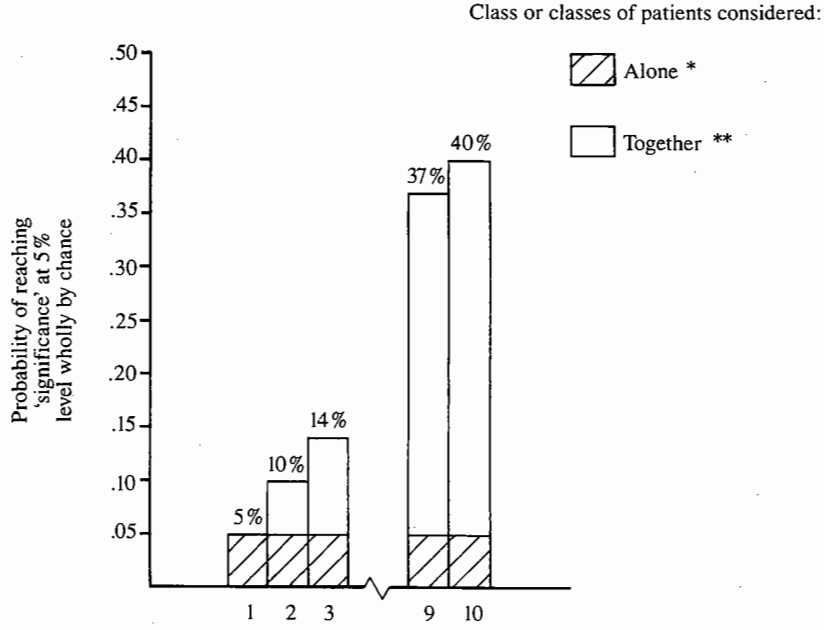
One thing is clear, Sterling observed: the risk of Type I error stated by the author should not be accepted at its face value. Since the general reluctance to publish negative results is easily confirmed by reading current medical articles, the cautionary advice is quite relevant.

## 'Significant' results in sub-classes

I have emphasized that in a focused trial the class of patients and an end point to be considered are clearly specified in advance; the quality of evidence is strengthened by limiting the problems associated with multiplicity.

The relationship between strength of evidence and multiplicity is described by Tukey in terms of the simple arithmetic of asking multiple ques-

**Multiplicity of classes of patients and the probability of a 'significant' outcome**




*For each class considered alone the probability of finding a 'significant' result is 100 per cent minus 95 per cent = 5 per cent

**For two classes the probability of finding at least one of the two results 'significant' is 100 per cent minus (95 per cent)$^2 \simeq 10$ per cent ... for ten classes the probability of finding at least one out of the 10 results 'significant' is 100 per cent minus (95 per cent)$^{10} \simeq 40$ per cent. (Based on compilation by Tukey)

tions and concentrating on the most favorable answers. In a hypothetical inquiry concerning the effect of a treatment that is perfectly neutral (totally without effect on any patient), as the classes of patients examined is increased, the likelihood of finding a 'significant' result in at least one of the subgroups purely by chance rises sharply. Thus, the evidence that for one class of patient, a broad inquiry has reached some specified level of 'statistical significance,' is much less powerful than the finding that pre-study specification of one class has reached exactly the same level of 'significance'.

Obviously, it is right for a physician to *want* to know about the behaviour to be expected from a therapy when applied to his particular patient. But, Tukey points out, he cannot *expect* this. The national RLF study clearly indicated that such expectations are unrealistic. The limits of the generalizations that could be drawn from this time-consuming effort were quite narrow: RLF outcome was examined in 75 subclasses of patients, but questions concerning these 'passive' associations could only be resolved by further testing.

## CONCEPTS OF STATISTICAL PROOF

### 'Significance' level as direct probability

A restriction in the concept of statistical proof can be seen by examining a distinction between finding an effect when we are given the causes and finding a cause when we are given effects. Statistician Richard B. Darlington of Cornell University points out that in the first situation we can make a prediction by calculating what is called 'direct' probability; for example, in coin tossing we compute the chances of 'heads' knowing the circumstances under which they are to occur. The second situation is an inverted one often found in medical experiments. We observe the event and ask, What is the probability that results from its occurrence in favor of any set of circumstances under which the same might have happened? The rules for calculating the latter so-called 'inverse' probability cannot be formulated without more information about the distribution of *causes*.

For instance, at the beginning of the national RLF trial, it was postulated (as in classical hypothesis testing) that oxygen curtailment would *not* reduce the risk of RLF by a specified amount. And it was necessary to *assume* that this hypothesis was true in order to calculate the probability of occurrence of the disorder in a group of babies treated in the new way. The no-important-reduction postulate was rejected because the observed reduction in the occurrence of blindness satisfied a pre-set 'significance' level (that is, *if* the hypothesis was, in fact, true, a reduction in RLF rate from 23 per cent to 7 per cent would be expected in only one of more than 100 trials of similar size). The reverse operation—calculating the probability that the



no-important-reduction hypothesis was true on the basis of the actual occurrence of RLF under curtailed oxygen—could not be justified.

These convoluted sentences make the subtle but important distinction between forward and backward inference. A 'significance level' is a kind of direct probability: the probability of observing a certain type of event *if* the null hypothesis is true. Inverse probability concerns the probability that the null hypothesis is true if an event is observed. Darlington reminds us that in a purely scientific inquiry, we would really like to know the inverse probability, but we must normally be satisfied with knowing only the direct probability. To calculate the former, we must have some estimate of the probability of the hypothesis *before* knowing the data; this is called prior probability of the hypothesis. It is used in Bayesian analysis, a statistical approach based on a theorem proposed by Reverend Thomas Bayes, an eighteenth-century minister.

**The Reverend Thomas Bayes' approach to inferences concerning hypotheses**

('Subjective' or 'personal' probability)

What can be inferred about the probability of concurrence of a theorized cause and an outcome of interest from observations of associations between the two? For example, based on the observations in Melbourne hospitals reported in 1951 (p 19) concerning the co-incidence of 'free use of oxygen' and RLF, what is the probability of finding this association in *future* observations?

| | RLF | | |
|---|---|---|---|
| *Oxygen Use* | *Present* | *Absent* | *Total* |
| Free use | $a$ | $b$ | $a+b$ |
| | 23 | 100 | 123 |
| Conservative use | $c$ | $d$ | $c+d$ |
| | 4 | 54 | 58 |
| Total | $a+c$ | $b+d$ | $a+b+c+d$ |
| | 27 | 154 | 191 |

(1) The relative frequency (or 'probability') of occurrence of the outcome of interest among those exposed to the theorized cause was

$$p=\frac{a}{a+b}=\frac{23}{123}=0.19$$

This represents a 'conditional probability': the probability of RLF *on condition* of a history of 'free use of oxygen'. (Similarly, the probability of RLF on condition of a history of 'conservative use of oxygen' was $4\div 58=0.07$)

(2) The relative frequency (or 'probability') of the theorized cause among those with the outcome of interest was

$$p=\frac{a}{a+c}=\frac{23}{27}=0.85$$

The probability of a history of 'free use of oxygen' *on condition* that RLF is present.

Reverend Bayes proposed a theorem as a basis for inductive inference that has led to an approach often termed 'Bayesian statistics'. In the application of this method, Mainland explained, we seek to arrive at the conditional probability stated in (2) from the probabilities (relative frequencies) alone. For example,

(3) As noted in (1), the probability of RLF on condition of a history of 'free use of oxygen' was 0.19

(4) But those exposed to the 'free use of oxygen' constituted a given proportion of the total

$$p=\frac{a+b}{a+b+c+d}=\frac{123}{191}=0.64$$

(5) Therefore, those with a history of 'free use of oxygen' who had RLF constituted

$$(3)\times(4)=0.19\times 0.64=0.12 \text{ of the total}$$

(6) All of the infants with RLF constituted

$$\frac{a+c}{a+b+c+d}=\frac{27}{191}=0.14 \text{ of the total}$$

(7) As noted in (5), 0.12 of the total were exposed-affected, thus,

$$p=\frac{(5)}{(6)}=\frac{0.12}{0.14}=0.85$$

The probability of finding a history of 'free use of oxygen' on the condition that RLF is present—the value obtained in (2).

(8) Bayes' Theorem takes this form to estimate the probability of concurrence of a theorized cause and an outcome of interest from observations of the two,

$$p=\frac{(1)\times(4)}{(6)}=\frac{0.19\times 0.64}{0.14}=0.85$$

On the assumption, belief, and hope that the observations on which these calculations are based represent a fair sample of 'all' infants at risk.

The 'free use of oxygen' is the hypothesis and RLF is a piece of data determined by observation. Note that the probability of 'free use of oxygen' was used in (4) as a relative frequency. It is what we would declare about an infant whom we knew to have a history of either 'free use of oxygen' or 'conservative use', if we knew the relative frequencies of these two histories, but *before* we knew about the presence or absence of RLF. It is therefore called a 'prior' or 'initial' probability in the classical approach. In Bayesian statistics, this is termed the 'prior probability of the hypothesis'—and in this sense it is a 'personal' probability, a statement of betting odds. (In fact, Bayes suggested that probability judgments based on mere hunches should be combined with probabilities based on relative frequencies.) Moreover, the conditional probabilities in (1), (6), and (8) are also considered 'personal' probabilities, in the sense that the frequencies and proportions favorable to total possibilities help form orderly and consistent opinions, rather than just any opinion. Statistical inference is modification of these opinions in the light of evidence, and Bayes' Theorem specifies how much modification should be made.

(Based on Mainland's exposition)



### 'Significance' as change in belief

The results of a trial with pragmatic emphasis also have some explanatory implications and our interest often extends beyond the purposes to which the results will be put. How can we describe the non-utilitarian aspect of conclusions drawn from a 'significance' level in some unambiguous way? Darlington suggests that the clearest interpretation is not as a *state* of belief but as a *change* in belief in a hypothesis. Words that imply this change are 'strengthen' or 'weaken'.

A result 'significant' at the 5 per cent level might be said to 'strengthen the alternative hypothesis' or to 'weaken the null hypothesis', depending on the interpreter's previous views. However, if an observer has no opinion before seeing the result, the outcome will produce a large change of personal view. In these circumstances, a term implying a state rather than a change of belief may be appropriate: 'the result confirms the experimental hypothesis at the 5 per cent level of significance' (and the antonym here is 'disconfirm').

### 'Not proven' verdict

The words used to describe 'non-significance' also need to be chosen carefully to avoid misunderstanding. R.W. Smithells of the University of Leeds has noted that English law insists that a prisoner is innocent until proven guilty. But the declaration at the end of a negative trial is less likely to mislead, he advised, when the rule of Scottish law is followed: a verdict of 'not proven' is permitted.

The latter terminology reminds us that failure to *demonstrate* differences does not prove that they are in fact negligible (p 119).

### Practical interpretation

At the end of a pragmatic trial we envisage some practical action that will be taken when a verdict statement is made. Even when we decide that there is no important difference between compared treatments, practical considerations may lead to a decision in favor of the innovation. Such action is not determined solely on the basis of 'significance level' since we prefer treatments that are readily accepted by patients, easy to administer, and inexpensive. The move, then, from decisions made about the results of a clinical trial to individual value judgments made at the bedside needs to be examined if we are to understand the process by which a medical innovation is translated into something of social value.

In the next chapter, I propose to discuss how individuals (the physician and the patient) and communities faced with a problem of choice under uncertainty may go about deciding on a course of action that is consistent with individual basic judgements and preference.

I fear that I have spent so much time in this chapter on the 'ifs', 'ands', 'buts', and 'howevers' when making statistical inferences that I may have given a mistaken impression that an approach based on the science of chance is a very weak one. (Harry Truman once confessed that he was looking for a one-armed economist to provide him with a solution for the country's economic woes. All of his advisors hedged their remarks by saying, 'On the one hand ..., but on the other hand ...') But (if I can be permitted one last adversative conjunction) there is little need to sing the praises of the goddess of fortune. She is generous in her rewards to those who stick with the odds. It is possible to 'break the bank at Monte Carlo' by betting against the house and dismissing the laws of probability with a snap of the fingers, but don't bet your life on it.

> For we know in part, and we prophecy in part ...
> For now we see through a glass darkly; ...
> I *Corinthians* 9 and 12