Measuring degrees of confirmation

GEORGE N. SCHLESINGER

1. Most philosophers agree that scientific confirmation amounts to an increase in the probability of a hypothesis $h$ due to the evidence $e$. They agree that

- if $P(h|e)$ (i.e. $h$'s conditional probability given $e$) > $P(h)$ (i.e. the prior probability of $h$), then $e$ confirms $h$;
- if $P(h|e) = P(h)$, then $e$ is irrelevant to $h$;
- if $P(h|e) < P(h)$, then $e$ disconfirms $h$.

However, conflicting opinions exist concerning what precisely represents numerically the degree of evidential support a hypothesis has received. In other words, there are various opinions as to how the growth of the credibility of a hypothesis, owing to the emergence of substantiating evidence, is to be evaluated. Recently the disagreement has narrowed down to two contestants:

- **Criterion D.** The degree of confirmation $e$ confers on $h$ equals the difference between the final probability of $h$ owing to the evidence and the prior probability – that is, $D(e, h) = P(h|e) - P(h)$.
- **Criterion R.** The degree of confirmation is measured by the ratio between $h$'s posterior and prior probability – that is, $R(e, h) = P(h|e)/P(h)$.

2. One argument in favour of Criterion R involves the case of three students, $a$, $b$, $c$, who are given a multiple choice test in which one and only one of the five printed answers is correct. Each student has to mark the answer he thinks correct. Let

- A: a knows the right answer,
- B: b knows the right answer,
- C: c knows the right answer;
- $T_i$: student $i$ marked the right answer.

There is strong past evidence that

- $P(A) = 3/5$, $P(B) = 2/5$, $P(C) = 1/5$.

We note that $P(T_a | A) = 1$. Also that $P(T_a | -A) = 1/5$. Furthermore, if $A$ does not know the answer, then we ascribe the same probability to his marking any one of the five printed answers. Consequently by Bayes's theorem,
MEASURING DEGREES OF CONFIRMATION

\[
P(A \mid T_a) = \frac{P(A)}{P(T_a \mid A)P(A) + P(T_a \mid \neg A)P(\neg A)} = \frac{3/5}{3/5 + (1/5 \times 2/5)} = \frac{15}{17}
\]

Thus \(R(T_a, A) = \frac{15}{17} - \frac{3}{5} = \frac{25}{42}\), and \(D(T_a, A) = \frac{25}{42} + \frac{25}{42} = \frac{25}{21}\).

Using the same method we obtain \(P(B \mid T_b) = \frac{10}{13}\), and thus \(R(T_b, B) = \frac{25}{13}\) and \(D(T_b, B) = \frac{24}{65}\). Finally, \(P(C \mid T_c) = \frac{5}{9}\); thus \(R(T_c, C) = \frac{25}{9}\), and \(D(T_c, C) = \frac{16}{45}\).

We note that \(R(T_c, C) > R(T_b, B) > R(T_a, A)\). So adopting Criterion \(R\) accords with the dictates of informal reasoning. Student \(c\) generates the biggest surprise by giving the correct answer. He was the least expected to do so, since the probability of his success was less than that of the other two. In view of the fact that \(c\)'s successful answer is the most remarkable, it has the greatest impact on our mind and thus is bound to generate a stronger pressure to raise the probability of \(C\) than e.g. a successful answer from \(b\) inclines us to raise the probability of \(B\). Similarly, the impact of observing \(T_b\) is greater than that of observing \(T_a\), and hence the credibility of \(B\) should rise by a greater percentage than that of \(A\).

On the other hand, \(D(T_c, C) < D(T_b, B)\), which puts Criterion \(D\) in conflict with our informal reasoning. Even more puzzling is to find Criterion \(D\) showing an inconsistency: it reverses the verdict when it comes to comparing correct answers from \(a\) and \(b\). For \(D(T_a, A) < D(T_b, B)\). So oddly, according to Criterion \(D\), the direction of increase in confirmation turns round between \(a\) and \(b\), and \(b\) and \(c\), when they all answer correctly.

3. The advocates of Criterion \(D\) think they have some valid arguments in support for their position. Donald Gillies is among the strong adherents of Criterion \(D\): he points out that in science, in general, \(P(e \mid h) = 1\) (with the exception of sciences based on statistical arguments). Thus in standard cases, Bayes's theorem yields \(P(e \mid h)/P(e) = 1/P(e)\), which means that by criterion \(R\), the support \(h\) receives from \(e\) is a function of \(P(e)\) and nothing else. This, he insists, is preposterous; surely the increase in the probability of \(h\) depends greatly on the prior probability of \(h\) as well! ([1], pp. 110–13).

Gillies is certainly right that the prior probability of \(h\) is bound to play a crucial role in determining the degree of confirmation. It is somewhat puzzling why he did not go even further, and decry the lack of reference to \(P(h \mid e)\), which too is an indispensable factor in deciding the degree to which \(e\) confirms \(h\). However, there is no room for any complaint; criterion \(R\) does not ignore the role of these factors, since if \(P(e \mid h) = 1\), then \(1/P(e) = P(h \mid e)/P(h)\), which shows precisely what we need – the degree of confirmation (i.e. the value of \(1/P(e)\)) is high if \(P(h)\) is small and that of \(P(h \mid e)\) is large.
In fact, our ability to represent the degree of confirmation by a simple expression like $1/P(e)$ is an advantage. For we may have indefinitely many hypotheses $h, k, l \ldots$ each implying $e$, without us knowing any of the values of $P(h), P(k), P(l) \ldots$ or any of the values of $P(h|e), P(k|e), P(l|e)$ etc., but if we knew the value of $P(e)$, we would have sufficient basis for asserting that evidence $e$ confirms each of these hypotheses to the same degree, $1/P(e)$. On the other hand, our knowledge of any single term relevant to $D(e, h)$ gives us the degree of confirmation of nothing else.

R. Rosenkrantz – like Gillies – also believes that he has an argument for adopting criterion D. He discusses a situation where $P(e|h) = 1$, and therefore $P(e|h \& k) = 1$, regardless what hypothesis $k$ stands for, as long as $P(k) > 0$. On the basis of R, $e$ provides equal degrees of confirmation for $h$ and $h \& k$, namely $1/P(e)$. But Rosenkrantz says

That seems objectionable when $k$ is extraneous, and even more objectionable when $k$ is probabilistically incompatible with $h$ in the sense that $P(k|h)$ is low. Personally, I have always considered this reason enough to reject the ratio measure in favor of a different measure:

$$D(e, h) = P(h|e) - P(h).$$

This measure is easily seen to satisfy the following condition

$$Z: D(e, h \& k) = P(k|h).$$

Subsequently, Rosenkrantz insists that $Z$ is the expression that delivers precisely what seems reasonable to us, namely that degree of confirmation $e$ confers to $h \& k$ is not independent of the relationship between $h$ and $k$, i.e. it varies with variations in the value of $P(k|h)$. That is, the smaller $P(k|h)$ is, the lower the confirmation that $e$ confers upon $h \& k$.

In fact, however, from the point of view of those who subscribe to Criterion R, what Rosenkrantz sees as a merit is by no means one. For according to them, both $R(e, h \& k)$ and $R(e, h)$ equal $1/P(e)$, and that is how it should be. What grounds are there for saying that the smaller $P(k|h)$ is, the lower the confirmation $e$ confers upon $h \& k$? In fact, whether $P(k|h)$ equals one or is almost zero makes no difference. Whatever the value of $P(k|h)$ it affects the prior probability $P(h \& k)$ – the denominator of $R$ – exactly the same way as it affects the final probability $P(k|h \& e)$. Thus the $P(k|h)$ in the numerator cancels that in the denominator.

4. I believe that the following is the strongest argument for adopting criterion R. Suppose the various commercial aircraft flying at present are all replaced by a newly developed plane which is faster, more comfortable, and less expensive to fly than any of the existing planes. In the opinion of
experts, it is a very safe vehicle, the probability of its crashing on a single flight being $1 \times 10^9$. All travellers would be eagerly looking forward to the opportunity to fly the new plane. However, to everybody's horror, from the very beginning each day several dozens of the epoch-making machines exploded soon after take off. Upon a meticulous inspection a basic structural flaw is discovered in the plane, in consequence of which in the estimate of experts there is a $1 \times 10^2$ chance that on a given flight any plane of the type is going to crash. Clearly, no reasonable person will entrust themselves to such a hazardous conveyance.

Now, during WWII, sometimes fighters landed gliders behind enemy lines. Only very skilled, strongly motivated and truly courageous people consented to participate in these high risk attacks. Quite likely, those few who did volunteer would not have done so if a fatal end was a certainty or even if the probability of violent death was 95% or 90%. Let us assume that by common knowledge the chances of perishing were 26%. In view of the common knowledge of the consequences of Nazi victory, it is conceivable that some airmen were ready to face the odds and volunteer to fly behind enemy lines. Suppose now that one day the commander announces that due to peculiar weather conditions the risk of losing one's life in a glider operation has increased from 26% to 27%. It may reasonably be assumed that anyone ready to face a 26% risk is unlikely to withdraw suddenly, owing to his paralysing fear when the risk increases to 27%.

According the Criterion R the degree to which the discovery of the structural flaw in the plane raised the probability, or provided confirmation for, a crash is $1/100$ divided by $1/10^9$, i.e. $10^7$. On the other hand, the extraordinary weather conditions confirmed a catastrophic end for the glider mission to a degree given by $27/100$ divided by $26/100$, i.e. $1.038$. There is an enormous difference between $10^7$ and $1.038$, and this well reflects the fundamental difference in people's reaction to the respective increased risk in the two stories. In the first case, there is a fundamental change from unquestioning trust in the soundness of the aircraft to complete loss of confidence. In the second case the changed risk is deemed far too trivial to make the slightest difference.

On the other hand, by Criterion D, observing the defective part of the plane confirmed a crash to a degree given by $1/100 - 1/10^9$, i.e. $0.00999...$, while the unusually bad weather confirmed a fatal danger involved in the gliding operation to a degree of $27/100 - 26/100$, i.e. 0.01, which is greater than $0.00999...$

Here we have an extreme example of a conflict between what many philosophers maintain qua philosophers and what sane people in general indicate through their actual behaviour to be the right way of evaluating degrees of confirmation.
Even if the adherents of Criterion D advanced what seem genuinely strong arguments in support of their position, it is most unlikely that people would change their convictions, especially when facing choices on which their life depended. They would continue to make their decision on the basis of Criterion R which has firm, intuitive support.

Admittedly, as many hold, philosophical analysis is not a mere servant of common sense but a superior tool in our search for the truth. It is one of the major tasks of philosophers to scrutinize meticulously commonly held views and revise or reject them when rigorous analysis finds them misguided. This is indeed so when the issue is one of interpretation, but not when it involves practical decisions. Given that in situations like those depicted in the story of the new aeroplanes and the story concerning the glider, virtually everyone, including philosophers, is going to behave as stated, the rational thing to do in response to those who believe they have compelling arguments in favour of Criterion D is to search for a flaw in their reasoning.

University of North Carolina, Chapel Hill, NC 27599-3125, USA

References