The Probability of the Evidence

Sherri Roush

Commentary to date on the term "P(e)" of the Bayes equation, $P(h/e) = P(e/h)P(h)/P(e)$, says, first, that P(e) should not be 1 because then $P(h/e) = P(h)$ and e cannot be evidence for h, because e fails to be probabilistically relevant to h, leading to the familiar problem of old evidence. Second, it says that a low value for P(e) makes sense of the intuitive idea that surprising evidence is more confirming than evidence which is not surprising. These points, together with the fact that P(e) is in the denominator of the right-hand side of the Bayes equation, suggest that it would be a bad thing for e's status as evidence for h if P(e) had a high value less than 1. This paper argues that for both technical and intuitive reasons this suggestion is false. As I show mathematically, high values for P(e) combined with high values for the likelihood ratio, $P(e/h)/P(e/-h)$, put a lower bound on the value of the posterior probability of the hypothesis, P(h/e). As I argue intuitively, a scheme in which we determine high values for P(e) and for the likelihood ratio in order to determine that e is evidence for h makes sense of the familiar practice of eliminative reasoning in science and elsewhere. And, as I argue, P(e) must be high to justify Bayesian conditionalization on e.

Commentary to date on the term "P(e)" of the Bayes equation,

$$P(h/e) = P(e/h)P(h)/P(e),$$

says, first, that P(e) should not be 1 because then $P(h/e) = P(h)$ and e cannot be evidence for h, because e fails to be probabilistically relevant to h.[1] (Glymour 1980, ch. 3) Second, it says that a low value for P(e) makes sense of the intuitive idea that surprising evidence is more confirming. (Howson and Urbach 1993, 123-126) These points, together with the fact that P(e) is in the denominator of the right-hand side of the Bayes equation, suggest that it would be a bad thing for e's status as evidence for h if P(e) had a high value less than 1. This paper argues that for both technical and intuitive reasons nothing could be further from the truth.

Even if one does not think, as I do, that the Likelihood Ratio, $P(e/h)/P(e/-h)$, is the one true measure of confirmation, any Bayesian will have to admit that it is a good thing

---

[1] P(e) = 1 alone implies that $P(h/e) = P(h)$ because P(e) is always between P(e/h) and P(e/-h) inclusive, meaning that either P(e/h) = 1 or P(e/-h) = 1 (or both). In the first case P(h) = 1, in the second P(-h) = 1.

and not a bad thing if P(e/h)/P(e/-h) is greater than 1. This condition says that e is more

likely if the hypothesis h is true than if it is false, and implies positive relevance. I will

assume in addition that the greater the Likelihood Ratio is, the better a thing it is. If we

suppose that the Likelihood Ratio (LR) is greater than 1, and follow the consequences as

it increases, then an interesting trend emerges: lower bounds on P(e) yield lower bounds

on P(h/e), the posterior probability of the hypothesis. This can be seen by graphing the

following rearrangement of the Bayes equation,

$$P(h/e) = (LR - P(e/h)/P(e))/(LR -1),^2$$

putting LR on the y-axis and P(e/h) on the x-axis, and assigning increasing values to P(e)

on separate sections.[3] It is easy to see from the equation that values for P(e) and for

P(e/h) and P(e/-h) are sufficient to determine the posterior probability of the hypothesis.

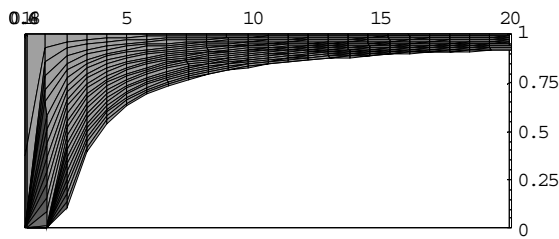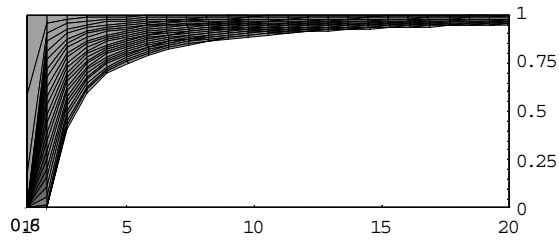Figure 1 shows how they do so when LR >1 and P(e) = .4.

Figure 1. P(e) = .4

Figure 2. P(e) = .5

---

[2] This equation can be derived by substituting (P(e)-P(e/-h))/(P(e/h)-P(e/-h)) in for P(h) in the standard Bayes equation. The equality of P(h) and the substitution can be derived by solving P(e) = P(e/h)P(h) + P(e/-h)P(-h) for P(h) and substituting 1 – P(h) for P(-h).
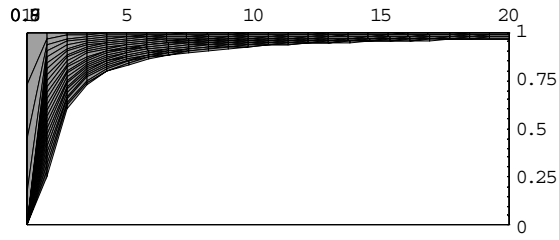[3] In other words, the equation becomes: P(h/e) = z = f(x,y) = (y-x/P(e))/(y-1).

Figure 3. P(e) = .6

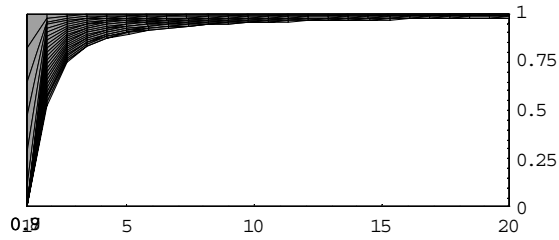

Figure 4. P(e) = .7
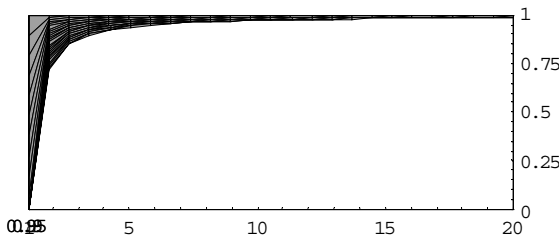
Figure 5. P(e) = .8

Figure 6. P(e) = .9
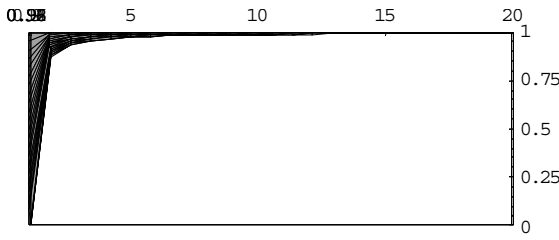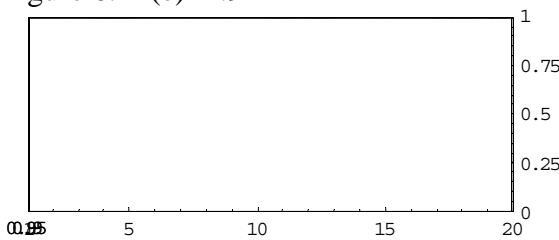

Figure 7. P(e) = 1

Figures 2 through 7 show what the plot looks like when LR > 1 and P(e) = .5, .6, .7, .8,

.9, and 1 respectively. The x axis is perpendicular to the page, its positive side starting at

the page and going straight in, which accounts for the garbled label numbers on the left

side—they are stacked in front of each other. It also means that the empty space under

3

the graphed surface on the right side shows that there are *no* values of $P(e/h)$ for which $P(h/e)$ is below the curve formed by the bottom of the visible graphed surface. Through graphing $P(e/h)$, $P(e)$ and the LR, we have found a result that is independent of $P(e/h)$ and depends only on $P(e)$ and the LR. For LR greater than about 3, fixed LR with increasing $P(e)$ yields increased minimum values for $P(h/e)$. Likewise, for LR greater than about 3, fixed $P(e)$ with increasing LR yields increased minimum values for $P(h/e)$. In general, putting a lower bound on the LR, and a lower bound on the value of $P(e)$ will put a strict lower bound on the posterior probability of the hypothesis. (See Appendix.)

A lower bound on the posterior probability of the hypothesis can only be good, because knowing this means knowing something about whether the hypothesis is true, which we do not get from positive relevance or even from high LR. If e is positively relevant to h, that is, $P(h/e) > P(h)$, this does not put a lower bound on the posterior probability of h unless we have a lower bound for $P(h)$. For, obviously, e may be positively relevant to h, that is, raise h's probability from what it was without taking e into account, while h started out with a low probability and e did not raise it enough to make h even more likely than not to be true. That is, e may be positively relevant to h while giving us no good reason to believe h. Likewise for the LR; the LR may be high while the posterior probability of the hypothesis is below .5. A high LR does not show that e gives us good reason to think h is true.

A condition on the notion of evidence that yields a lower bound for the posterior probability of the hypothesis is the only way to insure that having evidence for h means having some or good reason to believe h.[4] An obvious way to constrain the posterior

---

[4] I do not say that every notion of evidence must yield the consequence that evidence for h gives good reason to believe h. In fact, there are comparative notions that do not, e.g. evidence that gives you more

probability of the hypothesis is to put a lower bound on the prior probability of the hypothesis. If P(h) > .5 and P(h/e) > P(h), which is implied by LR > 1 as well as by positive relevance, then P(h/e) > .5. But while this condition is easy to announce, it is harder to know when a given case fulfills it. The condition whose consequences are shown in the graphs has the virtue of not requiring us to evaluate P(h) or P(-h) in order to determine P(h/e). Values of P(h) and P(-h) are of course involved in these graphs, but they are not evaluated. Rather, evaluations of P(e) and the likelihoods of e on h and on -h are sufficient to determine them, as can be seen from the equations:

P(e) = P(e/h)P(h) + P(e/-h)P(-h),

and

1 = P(h) + P(-h).

Given values for P(e) and the likelihoods P(e/h) and P(e/-h), we have two equations in two variables, and can solve for P(h) and P(-h) if we wish. P(h) and P(-h) are weights, reflecting how well each of the likelihoods, P(e/h) and P(e/-h), matches P(e). One may wonder whether it is any easier to evaluate P(e) than it is to evaluate P(h) or P(-h). I will argue below that it is frequently possible to evaluate P(e) directly.

It is obvious that there are many different combinations of high LR and high P(e) that will yield high P(h/e), and there are even many combinations of values that will lead to a particular high value for P(h/e). Moving from lower to higher, the first point where a P(e) value of .5 gets you a meaningful constraint on the posterior, P(h/e), is at roughly LR > 3, for there P(h/e) is >.5. That is, if it is more probable than not that e is true, and e is three times more likely on h than on -h, then it is more probable than not that h is true

---

reason to believe a hypothesis than you had before (positive relevance) and evidence that gives you more reason to believe one hypothesis than to believe another (high comparative likelihood ratio).

given e. This threshold is significant because it is easier to determine whether a statement is roughly more probable than not, that is, roughly greater than .5, than it is to assign an exact value to its probability. Also, if we know that the evidence makes the hypothesis more probable than not, that looks sufficient for counting us as having in e some reason to believe h.

Other points of interest occur as the LR and P(e) are pushed higher, since P(h/e) then approaches 1. Many and various combinations of values for the LR and P(e) will give a high or very high probability to P(h/e), and which probabilities are easiest to judge in a given situation should play a role in deciding which combinations of LR and P(e) thresholds to use. One can also work backwards from the threshold on P(h/e) that is good enough to yield good reason to believe, which may vary with circumstances, for example with the cost of being wrong. If in one's circumstances P(h/e) > .82 seems sufficient for good reason to believe, then finding P(e) > .75 and LR > 3 will be enough to give you what you want. If P(h/e) > .95 is desired, then it is useful to know that LR > 3 and P(e) > .92 will ensure it. If it is hard to get a read on your P(e), and you would like to show high P(h/e) by showing very high LR instead, you might choose to try to find that P(e) > .5 and LR > 21, because those together will insure that P(h/e) > .95.   Another combination of interest is LR > 7 and P(e) > .75, for these conditions together force the posterior P(h/e) to be greater than .95.

It is all very well for numbers to be related to numbers in a pleasant way, one might say, but what, one might ask, could be the rationale for requiring a high value of P(e) in real cases of evidence? The answer to this question comes in two parts, the first having to do with the term P(e) alone, the second with the compelling and familiar

picture of reasoning that emerges from combining a requirement of high P(e) with a requirement of high LR.

To see why it is good if P(e) is high in real cases, we must be careful about what the term "P(e)" represents. P(e) is often called the "expectedness" of the evidence, but this phrase can lead to misunderstanding. We get a better understanding if we approach the matter in the most straightforward way. On a personalist interpretation of probability, which I will assume for present purposes, P(s) is the degree of your belief in the statement s, and so, P(e) is the degree of your belief in e, according to the function P. There is no sense of temporal priority or expectation built in to P(e) that tells us under what circumstances to evaluate this term. Rather, evaluating it under different circumstances gives us different answers, and correspondingly different implications in the Bayes equation, because different probability functions are appropriate. For example, if we evaluate the probability of e after you have done a Bayesian conditionalization on e, then $P'(e) = 1$. If we evaluate the probability of e before conditionalization on e then P(e) will be less than 1. If we evaluate the probability of e after conditionalization on some statement that implies e, then $P''(e) = 1$, and if before then not necessarily.

The term "expectedness" can be misleading because I may believe e to some high degree even though I did not previously expect e to be true. P(e) measures actual degree of belief, not how much you expected at some prior stage that you would believe e at this stage. On the other hand, expectedness can affect my actual degree of belief in e. That e should have occurred may be so surprising that I am more willing to believe that my eyes are deceiving me than that e did actually occur. In such a case we might respond to witnessing that e occurs by saying "I don't believe it," and if that statement is true (as opposed to merely an expression of exasperation) then P(e) is low. However, in a case

7

where I have a very low degree of belief that e occurred, what reason could there be to think that e is evidence of anything for me?

It is similar in cases of surprising scientific evidence: no one takes a deeply surprising occurrence as evidence of anything until he satisfies himself that it did indeed occur, but then he has a high degree of belief that it occurred, and $P(e)$ must be high. Back-scattering of alpha particles from thin gold foil was deeply surprising to Ernest Rutherford, but he wouldn't have taken the statement, e_s, that back-scattering occurred at a certain rate, as evidence of the existence of a nucleus in the atom unless he had done some checking to reassure himself that e_s was true. On first report that the scattering had occurred, he might have been so surprised that he thought it more likely that his graduate student had blundered than that the report was true. In that case, and at that point, it hardly makes sense to think of the statement e_s as evidence for Rutherford of the existence of a nucleus. Once Rutherford had the experiment checked and repeated, e_s became evidence for him, but then his degree of belief in e_s was high, and thus so was $P(e)$.

I have suggested two distinct claims about a high value for $P(e)$, first that it is sufficient for e to be evidence for something, second that it is necessary for e to be evidence for anything. The first claim says that high $P(e)$ is a good thing for evidence, the second that it is the only way for e to be evidence. The first claim is (almost) implied by the principle of Bayesian conditionalization itself, which says:

> When your degree of belief in e goes to 1, but no stronger proposition also acquires probability 1, set $P'(a) = P(a/e)$ for all a in the domain of P, where P is your probability function immediately prior to the change. (Howson and Urbach 1993, 99)

This principle says that your degree of belief in e approaching 1 is a sufficient condition for conditionalizing on e, that is, for you to update your beliefs on the assumption that e is true (provided that no claim logically stronger than e also had its degree of belief approach 1). Your degree of belief in e prior to the conditionalization is just P(e), so high P(e) is (almost) sufficient for you to take e as evidence for whatever e happens to be positively relevant to, that is, to conditionalize upon it. Roughly, if you are confident of e, then you ought to let your other beliefs feel the appropriate effects of e's truth.

Bayesians often think of P(e) as your degree of belief in e before you observe e, on the assumption that you can observe e and thereby come to be certain of it, at which point you conditionalize upon it. A lot of evidence is not like that, e.g. Rutherford's *rate* of back-scattering is not something one comes to be certain of in one observation. But even in cases where we do come to be confident of e through one observation, P(e) remains by definition your degree of belief in e before you conditionalize upon e. If, as is often assumed, P(e) is your degree of belief as far back as before you observe e, then you have no reason to conditionalize upon e. It seems to me inescapable that in order for the value of P(e) that precedes Bayesian conditionalization to justify Bayesian conditionalization P(e) must be high.[5]

The claim that high P(e) is necessary for evidence is not implied by the principle of Bayesian conditionalization, but is implied by the claim that this should be the only rule of updating that one follows. Whether the principle and the claim that it is exclusive are right is obviously a controversial matter, as is whether they could be non-circularly

---

[5] Commentators are frequently misled by the fact that P(e) is in the denominator of Bayes's equation to conclude that P(e) is inversely proportional to P(h/e). (Howson and Urbach 1993, 124) This suggests that low P(e) will encourage high P(h/e). But there is no inverse proportionality here, because there is no constant of proportionality, since the other terms that stand between these two are variables that are not independent of P(e) or each other.

justified even if they were right. However, Bayesians take the first to be at least roughly

right, and we can motivate the second claim rather simply: when you do a Bayesian

conditionalization on e, you are effectively assigning probability 1 to e, that is, assigning

it the status of having the maximum possible degree of belief. Why would you do that if

you did not have a high degree of belief in e? This corresponds to the question why

Rutherford would take the claim that back-scattering occurred at a certain rate as

evidence of anything if he were not fairly confident that there was indeed back-scattering

from the gold foil at that rate.[6]

If we have a high value for P(e) and a high value for the Likelihood Ratio

P(e/h)/P(e/-h), then we are assured of a high posterior probability for the hypothesis.

What would it look like if we took this regularity as a guide for how to determine that

something is good evidence for a hypothesis? It would look like the eliminative

reasoning that scientists engage in on a daily basis. e appears very likely to be the case.

(Whether we make e occur, as in an experiment, or it occurs on its own, as in a

phenomenon not associated with an experiment, makes no difference to the current

point.) We have one or more hypotheses, h_1 through h_n, about what could be responsible

for e's occurrence, and there is also the possibility that what is responsible is something

we have not thought of, which can be represented as $-(h_1 \lor h_2 \lor h_3 \lor \ldots \lor h_n)$. We go

about using everything we know and can discover to rule out all of those possibilities that

we can. Sometimes this goes by means of background knowledge. For example, the fact

that DNA is a double-stranded ladder rules out some possible hypotheses about the way

---

[6] Of course, this only shows that no updating in which e was assigned probability 1 should be anything but
Bayesian. It does not rule out other forms of updating such as Jeffrey Conditionalization in which e is
assigned a probability less than 1. Appropriately more general things would need to be said to motivate a
more general version of the current scheme that allowed Jeffrey conditionalization and not only the

in which this molecule replicates. Those hypotheses will not even be candidates when it comes to inferring how DNA replicates from what we see in lab experiments. Sometimes this goes by measurements. For example, a disturbing magnetic field could not be responsible for the phenomenon because the magnetometer tells us there is no field in the vicinity. Sometimes, hypotheses can be ruled out because of our confidence in the way an experiment has been designed and executed. For example, body mass could not be responsible for the effect we see in the mice, because we randomized the sample for body mass. Ideally, only one possibility is left, and that is the hypothesis whose truth is responsible for the occurrence of e.

A hypothesis can be ruled out either because we know it is exceedingly unlikely or because even if it were true that would not make e likely. Such findings correspond to knowing that $P(h_m)$ is very low or that $P(e/h_m)$ is very low. Notice that findings of this sort are what we need in order to evaluate the likelihood of e on all of the possible hypotheses. $P(e/h_1)$ is the probability that e is true given h_1, and $P(e/-h_1)$ is equal to the following:

$P(e/h_2)P(h_2/-h_1) + P(e/h_3)P(h_3/-h_1) + \ldots + P(e/h_n)P(h_n/-h_1) + P(e/-(h_2 \vee \ldots \vee h_n)P(-(h_2 \vee \ldots \vee h_n)/-h_1)$.

Each term of this sum contains a term for the probability of an alternative hypothesis on the assumption that the first, h_1, is not true, and a term for evaluating how likely e would be if that hypothesis were true, the very things we take into account when we try to determine what is responsible for the occurrence of e, and, accordingly, what e is

Bayesian variety. Tentatively, it would say that you should conditionalize on a greater probability for e whenever your degree of belief in e increased.

evidence for. Thus, evaluation of P(e/h)/P(e/-h) corresponds to the eliminative reasoning scientists engage in when they try to determine what e is evidence for.

Finding P(e) to be high, in this scheme, is just finding good reason to believe that e did or does in fact occur, a necessary precursor to any rational effort to "explain" the occurrence of e. Thus it is clear why evaluation of P(e) is often a straightforward affair, unlike evaluation of P(h) and P(-h). For e_s to be evidence for Rutherford that the atom has a nucleus he has to have good reason to believe that it is true that alpha particles back-scatter at a certain rate. But there is no difficulty imagining him assembling that good reason to believe, by repeating the experiment, and checking that every part of it is working as assumed. The same is true with all other experimental evidence. It is good if the scientist can assign P(e) = .99, and we generally know how to put ourselves in a position to do that even if we fail to do it in a particular case.

High (non-unitary) values for P(e) are good. Together with high values for P(e/h)/P(e/-h) they put a lower bound on the posterior probability of the hypothesis. Evaluating these two terms corresponds to familiar eliminative reasoning to determine what is responsible for e and consequently what e is evidence for. We do not have to evaluate the terms P(h) and P(-h) to use this scheme as long as we can evaluate P(e), which there is every reason to think that we can and do.

Appendix

Why does high P(e) combined with high LR put a lower bound on P(h/e)? Here is one way to explain it. First, consider the following consequence of the axioms, which comes from substituting the right-hand side of this rearrangement of the Bayes equation, P(e/h)P(h) = P(h/e)P(e),

in for the first term on the right-hand side of the equation of total probability for P(e):

P(e) = P(e/h)P(h) + P(e/-h)P(-h).

We get:

P(e) = P(h/e)P(e) + P(e/-h)P(-h),

which, solving for P(h/e), becomes:

P(h/e) = [P(e) – P(e/-h) P(-h)]/P(e).[7]

This equation is interesting, because the right-hand side takes the form:

(P(e) – X)/P(e).

This expression takes its highest value when X is 0, because X is dragging the expression away from taking its highest possible value, P(e)/P(e) = 1. Therefore, P(h/e) will be maximized when the term X is minimized.

This alone already suggests qualitatively why raising P(e) with high LR will raise P(h/e): the larger P(e) is, the less a fixed value of X will drag down (P(e)-X)/P(e). Of course, X is not necessarily fixed when P(e) goes up, because the terms P(e/-h)P(-h) and P(e) are not independent. However, in the conditions that I have discussed for evidence, in which we put lower bounds on both P(e) and LR, special things happen.

Consider what we can deduce if P(e) > .5 and LR > 3. From LR > 3 it follows that the maximum value for P(e/-h) is .33. This is because the maximum possible value for P(e/h) is 1, and LR >3 says that P(e/-h) must be no more than one third of P(e/h). These conditions also impose an upper bound on P(-h), which we can see as follows. Note that because e is the weighted average of P(e/h) and P(e/-h), that is,

P(e) = P(e/h)P(h) + P(e/-h)P(-h),

---

[7] I owe James Pearson thanks for discovering this equation.

P(e) will always be between P(e/h) and P(e/-h) inclusive. In particular, for any value for the LR that is >1, the ordering will be P(e/h) ≥ P(e) ≥ P(e/-h). Under our conditions, and with P(e/-h) at its maximum, .33, the highest value for P(-h) will occur when the LR is as low as possible, namely 3, because a higher LR would force a lower P(e/-h) than the maximum .33. The change that raising the LR produced would have to go to reducing P(e/-h) since P(e/h) cannot go higher than 1, the value it must have if P(e/-h) = .33 and LR is as much as 3. For maximizing P(-h), P(e) is best minimized to .5, because as P(e) gets higher it moves away from the P(e/-h) of .33, and thus lowers p(-h), which is the weight on P(e/-h) in the average that makes up P(e). With these values that will yield the maximimum value P(-h) can have, we solve the following simultaneous equations:

.5 =   P(h) + .33P(-h)
.33 = .33P(h) + .33P(-h)

to yield

P(h) = .25, P(-h) = .75.

The first of the simultaneous equations is an instance of

P(e) = P(e/h)P(h) + P(e/-h)P(-h),

and the second is a form of the axiom that says that the probabilities of a set of statements that exhaust the logical space must sum to 1:

1 = P(h) + P(-h).

This shows that the maximum value for P(-h) under our conditions for evidence is .75, and we showed above that the maximum value for P(e/-h) is .33. The term X in the equation above is the product of these two, and so has maximum value .25. Thus, by the equation above,

P(h/e) = (P(e) – P(e/-h)P(-h))/P(e),

when P(e) is greater than .5 (and LR is greater than 3), P(e/-h)P(-h) is less than or equal to .25, so P(h/e) is greater than or equal to (.5-.25)/.5 = .5.

The fact that P(e) is the weighted average of the two likelihoods we use, P(e/h) and P(e/-h), that is, that P(e) is between these two likelihoods, and the fact that the hypotheses we consider exhaust the logical space, yielding the equation,

P(h) + P(-h) = 1,

played crucial roles in this derivation. The derivation does not go through if we imagine instead that we are dealing with a comparative likelihood ratio, $P(e/h_1)/P(e/h_2)$, where h_1 and h_2 do not exhaust the logical space. Accordingly, a high comparative likelihood ratio does not show that we have good reason to believe the hypothesis in the numerator.

References

Glymour, Clark (1980). *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Howson, Colin, and Peter Urbach (1993). *Scientific Reasoning: The Bayesian Approach*. Second Edition. Chicago and La Salle, Ill.: Open Court Publishing.