Measuring Confirmation

Peter Milne

Department of Philosophy
University of Stirling
peter.milne@stir.ac.uk

7th Annual Formal Epistemology Workshop
Konstanz, September 2nd - 4th, 2010

The Bayesian picture: we have, for hypothesis h and evidence e relative to background information b,

$$\begin{cases} \text{confirmation when } P(h|eb) > P(h|b); \\ \text{disconfirmation when } P(h|eb) < P(h|b); \\ \text{neither confirmation nor disconfirmation when } P(h|eb) = P(h|b). \end{cases}$$

(I'm going to take for granted that $P(e|b) > 0$ and $P(b) > 0$.)

Incremental confirmation: e confirms h just in case e increases degree of belief in h

This is all I'm interested in here. Jim Joyce has distinguished various conceptions of confirmation. Certainly there are other notions relating to the bearing of evidence on hypotheses and *vice versa*. It's contestable that there are other Bayesian conceptions of confirmation.

— Incremental confirmation, the *additional* evidence provided by e regarding h, a.k.a. *probative value* ...and how to measure it.

From an example of David Christensen's:

> *Wondering whether there are deer in a nearby wood, I come across first deer droppings then a discarded antler. Both are in themselves strong evidence for the presence of deer, but having come across the deer droppings, the antler provides little additional evidence.*

The Bayesian considers confirmation and disconfirmation to relate to what is *added* by a proposition over and above the support that background knowledge already provides.

The Bayesian is *not* providing an explication of the relation **e** *is good evidence for* **h**. In the ordinary run of things, both deer droppings and a discarded antler are good evidence for the presence of deer. But in a given epistemic context only one, or maybe neither, may serve to raise degree of belief significantly.

Confirming evidence increases degree of belief, disconfirming decreases it.
How do we measure this change?
There are three standard ways to quantify change:

- Difference: $P(h|eb) - P(h|b)$;
- Ratio: $P(h|eb)/P(h|b)$
- Proportional difference: $\dfrac{P(h|eb) - P(h|b)}{P(h|b)}$

But degrees of belief are *equally well* represented by *odds*.
The fundamental problem:

- sameness of difference in probabilities does not track sameness of difference in odds (and *vice versa*)
- sameness of ratio of probabilities does not track sameness of ratio of odds (and *vice versa*)
- sameness of proportional difference in probabilities does not track sameness of proportional difference in odds (and *vice versa*)

—We have to look beyond standard measures of change. We do that by achieving a balance between *plausible principles* and *examples*.

1. $$\begin{cases} \text{confirmation when } P(h|eb) > P(h|b); \\ \text{disconfirmation when } P(h|eb) < P(h|b); \\ \text{neither confirmation nor disconfirmation when} \\ \quad P(h|eb) = P(h|b). \end{cases}$$

2. The probabilities that matter for the definition of $Conf_b(h, e)$, whether taken straight or used to determine odds, are some among the values $P(\cdot|b)$ and $P(\cdot| \cdot b)$ take on the sixteen truth-functional combinations of e and h.

3. $Conf_b(h, e) > Conf_b(h, f)$ when $P(h|eb) > P(h|fb)$.

4. *Wondering whether there are deer in the nearby wood, I and a companion come across deer droppings. We agree that this is strong evidence for the presence of deer. We then come across an antler. We agree that, given the droppings, this doesn't add much support. I say, 'The droppings and the antler aren't much better evidence than just the droppings.'*
'No, on the contrary, taken together they're considerably better evidence,' she replies.
'But you agreed that the droppings are strong evidence and that the antler doesn't add much given the droppings.'
'Yes. That's right.'
'???'

At a loss to know how to make sense of the disagreement with my companion, I conclude

$Conf_b(h, ef)$ is determined by $Conf_b(h, e)$ and $Conf_{be}(h, f)$.

Notice that in the second term e is now part of the background.

The fourth principle,

$Conf_b(h, ef)$ is determined by $Conf_b(h, e)$ and $Conf_{be}(h, f)$

was called "generalized additivity" by I.J. Good.

Nozick's measure

$$n_b(h, e) = P(e|hb) - P(e|\bar{h}b)$$

and Christensen's measure

$$S_b(h, e) = P(h|eb) - P(h|\bar{e}b),$$

fail to satisfy constraint (3). [Fitelson]

Nozick's measure and Christensen's also fail generalized additivity (4).

Theorem Any measure of confirmation satisfying constraints (1), (2), and (4) can be expressed as a function of $P(h|b)$ and the ratio $\dfrac{P(h|eb)}{P(h|b)}$.

In addition to Nozick's and Christensen's measures, Carnap's

$\mathfrak{r}_b(h,e) = P(he|b) - P(h|b)P(e|b) = P(he|b).P(\bar{h}\bar{e}|b) - P(\bar{h}e|b).P(h\bar{e}|b)$,

Rescher's

$$\begin{aligned} Re_b(h,e) &= \frac{P(h|eb) - P(h|b)}{1 - P(h|b)} P(e|b), \text{when } P(h|eb) \geq p(h|b), \\ &= \frac{P(h|eb) - P(h|b)}{P(h|b)} P(e|b), \text{when } P(h|eb) < p(h|b), \end{aligned}$$

Crupi, Tentori, and Gonzalez'

$$\begin{aligned} Z_b(h,e) &= \frac{P(h|eb) - P(h|b)}{1 - P(h|b)}, \text{when } P(h|eb) \geq P(h|b), \\ &= \frac{P(h|eb) - P(h|b)}{P(h|b)}, \text{when } P(h|eb) < P(h|b), \end{aligned}$$

and the Odds Ratio

$$\frac{O(h|eb)}{O(h|\bar{e}b)} = \frac{P(he|b).P(\bar{h}\bar{e}|b)}{P(\bar{h}e|b).P(h\bar{e}|b)},$$

a measure of correlation widely used in medical statistics, all fail general additivity [Newcombe].

Corollary Up to multiplication by a positive constant, there is a *unique* non-trivial rescaling of any continuous measure of confirmation satisfying constraints (1), (2), (3), and (4) into a measure that scales confirmational neutrality as zero and adds across conjunctions of evidence: the extent to which *ef* confirms *h* relative to *b* is just the sum of the support *e* gives to *h* relative to *b* and the support *f* gives to *h* over and above that, *i.e*, to *h* relative to *be*.

An oddity of the ratio measure $R_b(h,e) = P(h|eb)/P(h|b)$

When *e* confirms *h* relative to *b*, $R_b(h,e)$ lies between 1 and ∞.
When *e* disconfirms *h* relative to *b*, $R_b(h,e)$ lies between 0 and 1.

On the face of it, confirmation always outweighs disconfirmation. But if it makes sense to speak of amounts of confirmation, something is badly amiss here. There is no natural sense in which confirming evidence always supports more than disconfirming evidence undermines.

If the net effect of the conjunction *ef* is neither to confirm nor disconfirm *h* relative to *b* then any confirmation (disconfirmation) due to *e* relative to *b* must be "undone" or "offset" by disconfirmation (confirmation) by *f* relative to *be*. If these amounts of confirmation and disconfirmation do not match up, there is an excess of confirmation over disconfirmation or *vice versa* that simply disappears.

The advocate of *R* has to say something along the lines that we should measure *disconfirmation* like this:

$$Disconf_b(h,e) = \frac{1}{Conf_b(h,e)} = \frac{1}{R_b(h,e)} = \frac{P(h|b)}{P(h|be)}.$$

But then, strictly speaking, confirmation and disconfirmation are being measured on different scales. They are related as density (mass per unit volume) and specific volume (volume per unit mass). But with good reason this is not, in methodological contexts, how we think of confirmation and disconfirmation. What we want is:
$Disconf_b(h,e) = -Conf_b(h,e)$.

To disconfirm *h* is to find evidence for its falsity, which is, *ipso facto*, to find evidence for the truth of $\bar{h}$. *I.e.* $Disconf_b(h,e) = Conf_b(\bar{h},e)$.

Hence

$$\textit{Hypothesis Symmetry}\text{: } Conf_b(\bar{h},e) = -Conf_b(h,e).$$

Numerous measures in the literature fail to satisfy Hypothesis Symmetry

- the ratio measure $R_b(h,e) = \dfrac{P(h|eb)}{P(h|b)}$
- the log ratio measure $r_b(h,e) = \log \dfrac{P(h|eb)}{P(h|b)}$
- the likelihood ratio measure $L_b(h,e) = \dfrac{P(e|hb)}{P(e|\bar{h}b)}$
- Haim Gaifman's $\dfrac{1-P(h|b)}{1-P(h|eb)}$
- Henry Finch's & Stephen Pollard's proportional difference measure $\dfrac{P(h|eb)-P(h|b)}{P(h|b)}$
- Lance Rip's $\dfrac{P(h|eb)-P(h|b)}{1-P(h|b)}$
- Popper's $\dfrac{P(h|eb)-P(h|b)}{P(h|eb)+P(h|b)} = \dfrac{P(e|hb)-P(e|b)}{P(e|hb)+P(e|b)}$
- Jim Joyce's $O(h|eb) - O(h|b) = P(\bar{h}|eb)^{-1} - P(\bar{h}|b)^{-1}$

A lot fail to meet the requirement of hypothesis symmetry. But we can rig up new measures, satisfying (1) - (4), that meet it. For example

$$\log \frac{\tan \frac{\pi}{2} P(h|eb)}{\tan \frac{\pi}{2} P(h|b)}.$$

And insisting that $Conf_b(\bar{h},e) = -Conf_b(h,e)$ doesn't force a measure of confirmation to be *strictly additive*:

$$Conf_b(h,ef) = Conf_b(h,e) + Conf_{be}(h,f).$$

The Kemeny–Oppenheim measure

$$ko_b(h,e) = \frac{P(e|hb)-P(e|\bar{h}b)}{P(e|hb)+P(e|\bar{h}b)}$$

is a (counter-) example.

What do measures like Christensen's $P(h|eb) - P(h|\bar{e}b)$ and the Odds Ratio $\dfrac{O(h|eb)}{O(h|\bar{e}b)}$ measure?

> *[Probative relations] compare the "posterior" evidence for h when e is added, to the "posterior" evidence for h when $\bar{e}$ is added. Here the issue is the extent to which the total evidence for h varies with changes in e's probability. When $P(h|eb)$ and $P(h|\bar{e}b)$ are close together, changes in $P(e|b)$ have little effect on $P(h|b)$, but when they are far apart such changes have a significant impact. [Hájek & Joyce]*

This is not *incremental confirmation*. What we have here are measures of how worthwhile it might be where h is concerned to find out *whether (or not)* e is the case. This is not the same as a measure of the impact on degree of belief in h of finding *that* e is the case.

What do measures like Nozick's $P(e|hb) - P(e|\bar{h}b)$, the Likelihood Ratio $\dfrac{P(e|hb)}{P(e|\bar{h}b)}$ and the Odds Ratio $\dfrac{O(e|hb)}{O(e|\bar{h}b)}$ measure?

> *These measures compare the "posterior" prediction of e when h is added, to the "posterior" prediction of e when $\bar{h}$ is added. Here the issue is the extent to which how much e is anticipated varies with changes in h's probability. When $P(e|hb)$ and $P(e|\bar{h}b)$ are close together, changes in $P(h|b)$ have little effect on $P(e|b)$, but when they are far apart such changes have a significant impact.*

Again, this is not *incremental confirmation*. What we have here are measures of the difference coming to believe h rather than $\bar{h}$ would make to one's degree of belief in e. This is not the same as a measure of the impact on degree of belief in h of finding *that* e is the case.

A card is drawn "randomly" from a deck of cards. e is 'It's the ten of diamonds'; h_1 is 'It's the ten of diamonds'; h_2 is 'It's the ten of diamonds or the six of clubs'; h_3 is 'It's the ten of diamonds or the six of clubs or the jack of diamonds'; ... and so on, in no particular order, through the entire pack, fifty-two hypotheses, progressively saying less and less until the last one, listing all the cards, is implied by background knowledge b.

Measures such as the difference measure $d_b(h,e) = P(h|eb) - P(h|b)$ and the ratio measure $R_b(h,e) = \dfrac{P(h|eb)}{P(h|b)}$ see the incremental confirmation going down, progressively, as we move through the sequence, to zero at h_{52}. For each $i, 1 \leq i \leq 52$, h_i receives the maximum confirmation possible *for it*, but since they say progressively less, these local maxima diminish as we go through the sequence.

Good's favoured log odds ratio measure $l_b(h,e) = \log \dfrac{P(e|hb)}{P(e|\bar{h}b)} = \log \dfrac{O(h|eb)}{O(h|b)}$ and the Kemeny–Oppenheim measure $ko_b(h,e) = \dfrac{P(e|hb) - P(e|\bar{h}b)}{P(e|hb) + P(e|\bar{h}b)}$ see matters quite differently.

They say that hypotheses h_1 to h_{51} all receive the same maximum confirmation: ∞ in the case of l, 1 in the case of ko. h_{52}, on the other hand receives no confirmation. (Constraint (1) tells us that there is neither confirmation nor disconfirmation of h_{52}.)

> *85% of taxis in a certain city are green, the rest are blue. An eye-witness to a hit-and-run accident testifies that the taxi involved was blue. On tests in similar conditions, she turns out to be 80% reliable in her judgments of taxi colour, by which we mean that on 80% of the occasions on which the object involved is blue, she reports it as being such, and likewise for occasions on which it is green.*

> *85% of taxis in a certain city are green, the rest are blue. An eye-witness to a hit-and-run accident testifies that the taxi involved was blue. On tests in similar conditions, she turns out to be 80% reliable in her judgments of taxi colour, by which we mean that on 80% of the occasions on which the object involved is blue, she reports it as being such, and likewise for occasions on which it is green.*

How does the amount of confirmation supplied by the witness's report vary with changes in the base rate and the witness's reliability?

Let $x = P(\textit{taxi is blue})$ and $y = P(\textit{witness says 'Blue'}|\textit{taxi is blue})$. Then

- difference measure: $d = \dfrac{x(1-x)(2y-1)}{xy + (1-x)(1-y)}$;
- log odds ratio measure: $l = \log y - \log(1-y)$;
- Kemeny–Oppenheim: $ko = 2y - 1$.

According to the log odds ratio and Kemeny–Oppenheim measures, *the base rate is irrelevant*. No matter how probable or improbable the involvement of a blue taxi, the witness's report affords the same amount of support to the hypothesis that the taxi involved in the accident was blue.

According to the difference measure, the lower the base rate the more inclined we are to write off the witness's report as mistaken, although the better her powers of colour discrimination the lower the base rate has to be in order to incline us significantly to do that. Conversely, if the percentage of blue taxis is large, we are unimpressed by her evidence, it's what we were expecting anyway. Somewhere in between lies the area in which we give most weight to the witness's report: it brings about most change in our degree of belief in the taxi's being blue.

Stuart Sutherland asserts:

> *In Great Britain about 300,000 people die each year from heart disease, while about 55,000 die from lung cancer. Heavy smoking approximately doubles one's chance of dying from heart disease, and increases the chance of dying from lung cancer by a factor of about ten.*

and goes on

> *Most people will conclude that smoking is more likely to cause lung cancer than heart disease and indeed both in Britain and elsewhere government campaigns against smoking have been largely based on this assumption. But it is false.*

In an obvious notation, Sutherland's figures tell us

$$\frac{P(h)}{P(l)} = \frac{300}{55};\ \frac{P(h|s)}{P(h|\bar{s})} = 2;\ \frac{P(l|s)}{P(l|\bar{s})} = 10.$$

We find that

$$\frac{P(h|\bar{s})}{P(l|\bar{s})} = \frac{P(h)}{P(l)} \times \frac{9P(s)+1}{P(s)+1}$$

and

$$\frac{P(h|s)}{P(l|s)} = \frac{2}{10} \times \frac{P(h)}{P(l)} \times \frac{9P(s)+1}{P(s)+1}.$$

As $\frac{300}{55} > 5$, $P(h|s) > P(l|s)$: irrespective of the proportion of the population that smokes, a smoker is more likely to die from heart disease than lung cancer.

Provided $0 < P(s) < 1$, the ratio, log-ratio, Popper and Finch/Pollard measures tell us tell us that smoking is better evidence of dying from lung cancer than from heart disease (and increasingly better evidence the smaller the percentage of the population that smokes).

The difference measure tells is that smoking is better evidence for death from heart disease unless the proportion of smokers in the population is less than $\frac{13}{147}$ (just under 9%).

Since ratios of odds are not determined by ratios of probabilities, the log odds ratio and Kemeny–Oppenheim measures tell us nothing regarding smoking as evidence.

(The measures all agree that death from lung cancer is better evidence for having been a smoker than is death from heart disease.)

Taking up Sutherland analysis of the statistics concerning deaths of smokers, we can estimate the proportion of deaths from heart disease in the general population were no-one to smoke by $P(h|\bar{s})$. Hence the number of deaths per unit population from heart disease attributable to smoking is given by $P(h) - P(h|\bar{s})$. Since

$$d(h,s) = P(h|s) - P(h) = (P(h) - P(h|\bar{s}))\frac{P(\bar{s})}{P(s)},$$

we have that more smokers 'kill themselves of heart disease [than] die from lung cancer caused by smoking' [Sutherland] if, and only if, smoking is better evidence of death from heart disease than death from lung cancer.

d factorizes as

$$(P(h|eb) - P(h|\bar{e}b))P(\bar{e}|b),$$

thus d takes account of the "probative force" of evidence e with respect to hypothesis h—the importance finding out whether e has for degree of belief in h—but weights it by the prior improbability of that evidence.

d factorizes as

$$(R_b(h,e) - 1)P(h|b) \text{ and } (1 - R_b(\bar{h},b))P(\bar{h}|b).$$

Both $R_b(h,e)$ and $R_b(\bar{h},e)^{-1}$ have been suggested as measures of severity of test. Going with that, d takes account of severity of test but weights it by the prior (im)probability of the hypothesis. Now, it may seem odd that when $P(e|h_1b) = P(e|h_2b)$ d says that the better confirmed of h_1 and h_2, if e confirms h_1 and h_2, or the more strongly disconfirmed, if e disconfirms h_1 and h_2 is the initially more probable, and, equally, it may seem odd that when $P(e|\overline{h_1}b) = P(e|\overline{h_2}b)$ d says that the better confirmed, if e confirms h_1 and h_2, or the more strongly disconfirmed, if e disconfirms h_1 and h_2 is the initially less probable. But ...

- if $P(e|h_1b) = P(e|h_2b) > P(e|b)$ then $P(h_1) > P(h_2)$ if, and only if, $P(e|\overline{h_1}b) < P(e|\overline{h_2}b)$;
- if $P(e|h_1b) = P(e|h_2b) < P(e|b)$ then $P(h_1) > P(h_2)$ if, and only if, $P(e|\overline{h_1}b) > P(e|\overline{h_2}b)$;
- if $P(e|\overline{h_1}b) = P(e|\overline{h_2}b) < P(e|b)$ then $P(h_1) < P(h_2)$ if, and only if, $P(e|h_1b) > P(e|h_2b)$;
- if $P(e|\overline{h_1}b) = P(e|\overline{h_2}b) > P(e|b)$ then $P(h_1) < P(h_2)$ if, and only if, $P(e|h_1b) < P(e|h_2b)$.

So, in the first case, for example, e better confirms h the less likely is e on the supposition that h is false. A symptom equally likely in the presence of two diseases is better evidence for the disease in whose absence it is less likely. Likewise, in the third case, a symptom equally (un)likely in the absence of two diseases, is better evidence for the disease in whose presence it is more likely.

So all in all I think the difference measure will do just fine as "the one true measure of confirmation".

- J. Martin Bland and Douglas G. Altman: 'The odds ratio', *British Medical Journal* **320** (2000), 1468.
- Rudolf Carnap: *Logical Foundations of Probability*, 2nd edn., Chicago: University of Chicago Press, 1962.
- David Christensen: Review of John Earman, *Bayes or Bust? A critical examination of Bayesian confirmation theory*, *Philosophical Review* **103**/2 (1994), 345-47.
- David Christensen: 'Measuring Confirmation', *Journal of Philosophy* **XCVI**/9 (1999), 437-61.
- Vincenzo Crupi, Katya Tentori, and Michel Gonzalez: 'On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues', *Philosophy of Science* **74** (2007), 229-52.

- Henry A. Finch: 'Confirming Power of Observations Metricized for Decisions among Hypotheses,' *Philosophy of Science* **27** (1960), 293-307, 391-404.
- Branden Fitelson: 'The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity', *Philosophy of Science* **66** (1999), S362-S378.
- Branden Fitelson: *Studies in Bayesian Confirmation Theory*, Ph.D. dissertation, Madison: University of Wisconsin–Madison, 2001.
- Richard D. Friedman: 'A Close Look at Probative Value', *Boston University Law Review* **66** (1986), 733-759.
- Haim Gaifman: 'Subjective Probability, Natural Predicates and Hempel's Ravens', *Erkenntnis* **14**2 (1979), 105-47.
- I.J. Good: 'Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments', *Journal of the Royal Statistical Society. Series B (Methodological)* **22**/2 (1960), 319 -31.

- Alan Hájek and James M. Joyce: 'Confirmation', in Stathis Psillos and Martin Curd (eds.), *Routledge Companion to the Philosophy of Science*, London: Routledge, 2008, pp. 115-29.
- James M. Joyce: 'Bayes' Theorem' in E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, `http://plato.stanford.edu/entries/bayes-theorem`. Accessed 14th August, 2010.
- Daniel Kahneman and Amos Tversky: 'On prediction and judgment', *Oregon Research Institute Bulletin* **12**/4 (1972).
- Daniel Kahneman and Amos Tversky: 'Evidential impact of base rates', in Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 1982, pp. 153-160.
- D. H. Kaye: 'Quantifying Probative Value', *Boston University Law Review* **66** (1986), 761-766.

- D. H. Kaye and Jonathan J. Koehler: 'The Misquantification of Probative Value', *Law and Human Behavior* **27**/6 (2003), 645-659.
- Theo A.F. Kuipers: 'The Success Theory of Confirmation, Part II: Quantitative Confirmation and its Qualitative Consequences', *Logique et Analyse* **42** (1999), 447-482.
- Deborah G. Mayo: 'Novel Evidence and Severe Tests', *Philosophy of Science* **58** (1991), 523-552.
- Robert G. Newcombe: 'A deficiency of the odds ratio as a measure of effect size', *Statistics in Medicine* **25** (2006), 4235-40.
- Robert Nozick: *Philosophical Explanations*, Cambridge MA: Harvard University Press, 1981.
- Fenna Poletiek: *Hypothesis Testing Behaviour*, Hove: Psychology Press, 2001.

- Fenna H. Poletiek: 'Popper's Severity of Test as an intuitive probabilistic model of hypothesis testing', *Behavioral and Brain Sciences* **32** (2009), 99-100.
- Stephen Pollard: 'Milne's measure of confirmation', *Analysis* **59**/4 (1999), 335-37.
- Karl R. Popper: *Conjectures and Refutations*, fourth edition, London: Routledge and Kegan Paul, 1972. (First edition, 1963.)
- Kameshwar Prasad, Roman Jaeschke, Peter Wyer, Sheri Keitz, Gordon Guyatt, and the Evidence-Based Medicine Teaching Tips Working Group: 'Tips for Teachers of Evidence-Based Medicine: Understanding Odds Ratios and Their Relationship to Risk Ratios', *Journal of General Internal Medicine* **23**/3 (2008), 635-40.
- Nicholas Rescher: 'A Theory of Evidence', *Philosophy of Science* **25** (1958), 83-94.

- Lance J. Rips: 'Two Kinds of Reasoning,' *Psychological Science* **12** (2001), 129-134.
- Stuart Sutherland: *Irrationality*, London: Constable, 1992. Republished as *Irrationality: the enemy within*, London: Penguin, 1994; reprinted under original title, London: Pinter & Martin, 2007.