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

Outline

1 Incremental confirmation
2 The fundamental problem
3 Basic Principles
   • Failures of general additivity
4 Disconfirmation and hypothesis symmetry
5 Measures of other things
6 Examples
   • The deck of cards
   • Kahneman & Tversky’s taxi-cab example
   • Sutherland’s deaths from smoking example
7 Some properties of the difference measure
8 Bibliography

Incremental confirmation

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

\[
\begin{align*}
\text{confirmation when } & P(h|eb) > P(h|b); \\
\text{disconfirmation when } & P(h|eb) < P(h|b); \\
\text{neither confirmation nor disconfirmation when } & P(h|eb) = P(h|b).
\end{align*}
\]

(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.
The fundamental problem

Confirmation increases degree of belief, disconfirmation 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: \( \frac{P(h|eb) - P(h|b)}{P(h|b)} \)

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.

The fourth principle,

\[
\text{Conf}_b(h, ef) \text{ is determined by Conf}_b(h, e) \text{ and Conf}_b(h, f)
\]

was called “generalized additivity” by I.J. Good.

Nozick’s measure

\[
n_b(h, e) = P(e|hb) - P(e|\overline{h}b)
\]

and Christensen’s measure

\[
S_b(h, e) = P(h|eb) - P(h|\overline{e}b),
\]

fail to satisfy constraint (3). [Fitelson]

Nozick’s measure and Christensen’s also fail generalized additivity (4).
An oddity of the ratio measure $R_b(h, e) = \frac{P(h|eb)}{P(h|b)}$.

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

$$R_{eb}(h, e) = \frac{P(h|eb) - P(h|b)}{1 - P(h|b)},$$

when $P(h|eb) \geq P(h|b)$,

$$= \frac{P(h|eb) - P(h|b)}{P(h|b)},$$

when $P(h|eb) < P(h|b)$,

Rescher’s

$$R_{eb}(h, e) = \frac{P(h|eb) - P(h|b)}{1 - P(h|b)},$$

when $P(h|eb) \geq P(h|b)$,

$$= \frac{P(h|eb) - P(h|b)}{P(h|b)},$$

when $P(h|eb) < P(h|b)$,

Crupi, Tentori, and Gonzalez’

$$Z_b(h, e) = \frac{P(h|eb) - P(h|b)}{1 - P(h|b)},$$

when $P(h|eb) \geq P(h|b)$,

$$= \frac{P(h|eb) - P(h|b)}{P(h|b)},$$

when $P(h|eb) < P(h|b)$,

An oddity of the ratio measure $R_b(h, e) = \frac{P(h|eb)}{P(h|b)}$.

When $e$ confirms $h$ relative to $b$, $R_b(h, e)$ lies between 1 and $\infty$.

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 $b$. 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.

and the Odds Ratio

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

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 $be$.

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 $\overline{h}$. I.e. $Disconf_b(h, e) = Conf_b(\overline{h}, e)$.

Hence

Hypothesis Symmetry: $Conf_b(\overline{h}, e) = -Conf_b(h, e)$. 
Numerous measures in the literature fail to satisfy Hypothesis Symmetry

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

What do measures like Christensen’s \( P(h|eb) - P(h|\bar{b}) \) and the Odds Ratio \( \frac{O(h|eb)}{O(h|\bar{b})} \) 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.

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 \( \text{Conf}_b(h, e) = -\text{Conf}_b(h, \bar{e}) \) doesn’t force a measure of confirmation to be strictly additive:

\[
\text{Conf}_b(h, ef) = \text{Conf}_b(h, e) + \text{Conf}_b(h, \bar{e}f).
\]

The Kemeny–Oppenheim measure

\[
k_{ob}(h, e) = \frac{P(e|hb) - P(e|\bar{b})}{P(e|hb) + P(e|\bar{b})}
\]

is a (counter-) example.

What do measures like Nozick’s \( P(e|hb) - P(e|\bar{b}) \), the Likelihood Ratio \( \frac{P(e|hb)}{P(e|\bar{b})} \) and the Odds Ratio \( \frac{O(e|hb)}{O(e|\bar{b})} \) 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 \( P(e|\bar{b}) \) 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 \(R_b(h, e) = \frac{P(h|eb)}{P(h|b)}\) 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.

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.

Good’s favoured log odds ratio measure

\[
l_b(h, e) = \log \frac{P(e|hb)}{P(e|\overline{hb})} = \log \frac{O(h|eb)}{O(h|\overline{b})} \text{ and the Kemeny–Oppenheim measure } k_{o_b}(h, e) = \frac{P(e|hb) - P(e|\overline{hb})}{P(e|hb) + P(e|\overline{hb})} \text{ see matters quite differently.}
\]

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

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 \(x = P(\text{taxi is blue})\) and \(y = P(\text{witness says ‘Blue’}|\text{taxi is blue})\). Then

- difference measure: \(d = \frac{x(1-x)(2y-1)}{xy+(1-x)(1-y)}\);
- log odds ratio measure: \(l = \log y - \log (1-y)\);
- Kemeny–Oppenheim: \(k_0 = 2y - 1\).
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

\[
\frac{P(h)}{P(l)} = \frac{300}{55}, \quad \frac{P(h|s)}{P(l|s)} = 2; \quad \frac{P(l|s)}{P(l|s)} = 10.
\]

Provided \(0 < P(s) < 1\), the ratio, log-ratio, Popper and Finch/Pollard measures 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 us 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 $P(h|\overline{s})$. Hence the number of deaths per unit population from heart disease attributable to smoking is given by $P(h) - P(h|\overline{s})$. Since

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

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

$$(P(h|e) - 1)P(h|b)$$ and $$(1 - R_b(h, e))P(h|b).$$

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.

### Measuring Confirmation

- if $P(e|h_1b) = P(e|h_2b) > P(e|b)$ then $P(h_1) > P(h_2)$ if, and only if, $P(e|h_1b) < P(e|h_2b)$;
- if $P(e|h_1b) = P(e|h_2b) < P(e|b)$ then $P(h_1) > P(h_2)$ if, and only if, $P(e|h_1b) > P(e|h_2b)$;
- if $P(e|\overline{h}_1b) = P(e|\overline{h}_2b) < P(e|b)$ then $P(h_1) < P(h_2)$ if, and only if, $P(e|\overline{h}_1b) > P(e|\overline{h}_2b)$;
- if $P(e|\overline{h}_1b) = P(e|\overline{h}_2b) > P(e|b)$ then $P(h_1) < P(h_2)$ if, and only if, $P(e|\overline{h}_1b) < P(e|\overline{h}_2b)$.

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”.


Haim Gaifman: ‘Subjective Probability, Natural Predicates and Hempel’s Ravens’, *Erkenntnis* 142 (1979), 105-47.


