"Weight of Evidence, Evidential Completeness and Accuracy"

Rafal Urbaniak and Marcello Di Bello

1 Motivations

1.1 Balance vs. weight

Suppose we want to represent our uncertainty about a proposition in terms of a single probability that we assign to it. It is not too difficult to inspire the intuition that this representation does not capture an important dimension of how our uncertainty connects with the evidence we have or have not obtained. In a 1872 manuscript of *The Fixation of Belief* (W3 295) C. S. Peirce gives an example meant to do exactly that.

> When we have drawn a thousand times, if about half have been white, we have great confidence in this result. We now feel pretty sure that, if we were to make a large number of bets upon the color of single beans drawn from the bag, we could approximately insure ourselves in the long run by betting each time upon the white, a confidence which would be entirely wanting if, instead of sampling the bag by 1000 drawings, we had done so by only two.

The objection is not too complicated. Your best estimate of the probability of $W =$'the next bean will be white' is .5 if half of the beans you have drawn randomly so far have been white, no matter whether you have drawn a thousand or only two of them. But this means that expressing your uncertainty about W by locutions such as "my confidence in W is .5' does not capture this intuitively important distinction.

Similar remarks can be found in Peirce's 1878 *Probability of Induction*. There, he also proposes to represent uncertainty by at least two numbers, the first depending on the inferred probability, and the second measuring the amount of knowledge obtained; as the latter, Peirce proposed to use some dispersion-related measure of error (but then suggested that an error of that estimate should also be estimated and so, so that ideally more numbers representing errors would be needed).

Peirce himself did not call this the weight of evidence (and in fact, used the phrase rather to refer to the balance of evidence, W3 294) [CITE KASSER 2015]. However, his criticism of such an oversimplified representation of uncertainty anticipated what came to be called weight of evidence by Keynes in his 1921 *A Treatise on Probability*:

> As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either increase or decrease, according as the new knowledge strengthens the unfavourable or the favourable evidence; but something seems to have increased in either case,—we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the weight of an argument. New evidence will sometimes decrease the probability of an argument but it will always increase its 'weight.' (p. 71)

The key point is the same [CITE LEVI 2001]: the balance of probability alone cannot characterize all important aspects of evidential appraisal. Keynes also considered measuring weight of evidence in terms of the variance of the posterior distribution of a certain parameter, but was quite attached to the idea that weight should increase with new information, even if the dispersion increase with new evidence [TP 80-82], and so he proposed only a very rough sketch of a positive sketch. Moreover, as he was uncertain how a measure of weight should be incorporated in further decision-making, the was skeptical about the practical significance of the notion. [TP 83]

But what is this positive sketch? On one hand, Keynes [TP 58-59] connects the notion of weight with relevance. Call evidence E relevant to X given K just in case $\mathsf{Pr}(X|K \wedge E) \neq \mathsf{Pr}(X|K)$.[1] One postulate than can be found in the *Treatise* [TP 84] is:[2]

(Monotonicity) If E is relevant to X given K, where K is background knowledge, $V(X|K \wedge E) > V(X|K)$, where V is the weight of evidence.

[RUNDE 1990, 280] suggests that Keynes at some point calls weight the completeness of information. This however, is a bit hasty, as Keynes only says that *the degree of completeness of the information on which a probability is based does seem to be relevant, as well as the actual magnitude of the probabiltiy, in making practical decisions*. As later on we will argue that it is actually useful to distinguish evidential weight (how much evidence do we have?) and evidential completeness (do we have all the evidence that we would expect in a given case?), we rather prefer to extract a more modest postulate:

(Completeness) If E_1 and E_2 are relevant items of evidence, and E_2 is (in a sense to be discussed) more complete than E_1, $V(X|K \wedge E_2) > V(X|K \wedge E_1)$.

If we conceptualize E_2 being complete and E_1 being incomplete as E_2 being a maximal relevant conjunction of relevant claims one of which is E_1, (Completeness) follows from (Monotonicity).

Similar requirements seem to be inspired by the urn example. We put them in two forms, a weaker and a stronger one.

(Weak increase) In cases analogous to the urn example, the weight obtained by a larger sample is higher, if the frequencies in the samples remain the same.

(Strong increase) In cases analogous to the urn example, the weight obtained by a larger sample is higher.

Now, some requirements on how weight of evidence is related to the balance of probability. For one thing, Keynes insists that new (relevant) evidence might decrease probability but will always increase weight [TP 77]. Since (Monotonicity) already captures the idea that weight will always increase, here we extract the other part of the claim:

(Possible decrease) It is possible that $V(X|K \wedge E) > V(X|K)$ while $\mathsf{P}(X|K \wedge E) < \mathsf{P}(X|K)$.

Clearly, Keynes also endorsed the following two requirements of a very similar form:

(Possible increase) It is possible that $V(X|K \wedge E) > V(X|K)$ while $\mathsf{P}(X|K \wedge E) > \mathsf{P}(X|K)$.

(Possibly no change) It is possible that $V(X|K \wedge E) > V(X|K)$ while $\mathsf{P}(X|K \wedge E) = \mathsf{P}(X|K)$.

Keynes is not referring to the sheer number of statements on the right hand side of a conditional probability P (H | E) or the sheer bulk of information that these statements contain. By "relevant evidence," Keynes is only referring to the extent that E pro- vides information that is pertinent to H in particular. [Pedden3]

1.2 Examples and informal desiderata

- Go over Nance in particular, Cohen, some other sources?

1.2.1 Monotonicity of weight

Runde, Joyce, Weatherson, Peden

1.3 Hamer's weight of evidence

1.4 Good's weigh of evidence and the information value

One notion in the vicinity also called *weight of evidence* has been introduced by Good [CITE PROBABILITY AND THE WEIGHING OF EVIDENCE 1950]. Let $W(H : E)$ be the Good's weigh of

[1] Keynes also uses a slightly more convoluted notion of relevance to avoid equally strong items of opposite evidence turning out to be irrelevant (this objection has also been brought up by [COHEN 1986 TWELVE]). The more complex version is that a proposition E_1 is relevant to X given K just in case it entails a proposition E_2 such that $\mathsf{P}(X|K \wedge E_2) \neq \mathsf{Pr}(X|K)$. [COHEN 1986 TWELVE] complaints that this still runs into difficulties. Ignore K, take an irrelevant proposition Z. It entails $Z \vee X$ and $\mathsf{P}(Z \vee X|X \wedge E) = 1$. Now, by Bayes' theorem we have $\mathsf{P}(X|E \wedge (Z \vee X)) = \frac{\mathsf{P}(X|E) \times \mathsf{P}(Z \vee X|X \wedge E)}{\mathsf{P}(Z \vee X|E)} = \frac{\mathsf{P}(X|E)}{\mathsf{P}(Z \vee X|E)}$. If the denominator differs from 1, the result differs from the numerator. We will ignore such difficulties, as they are not of key importance for the development of this chapter.

[2] RUNDE 1990 283 suggests Keynes allows for weight of evidence to decrease when new evidence increases the range of alternatives, but this is based on Keynes' claim that weight is increased when the number of alternatives is reduced, and Keynes does not directly say anything about the possibility of an increase of the number of alternatives.

evidence in favor of H provided by E (if we want to explicitly conditionalize on some background knowledge K, we write $W(H : E|K)$). One assumption about W taken by Good is as follows:

(Function) "It is natural to assume that $W(H : E)$ is some function of $\mathsf{P}(E|H)$ and of $\mathsf{P}(E|\neg H)$, say $f[\mathsf{P}(E|H), \mathsf{P}(E|\neg H)]$. I cannot see how anything can be relevant to the weight of evidence other than the probability of the evidence given guilt and the probability given innocence." [cite Good 1985 p 250]

The other two are:

(Independence) $\mathsf{P}(H|E)$ should depend only on the weight of evidence and on the prior: $\mathsf{P}(H|E) = g[W(H : e), \mathsf{P}(H)]$.

(Additivity) $W(H : E_1 \wedge E_2) = W(H : E_1) + W(H : E_2|E_1)$

The three conditions can be simultaneously satisfied by only one function (up to a constant factor), which leads to Good's definition of weight of evidence:[3]

$$W(H : E) = \log \frac{\mathsf{P}(E|H)}{\mathsf{P}(E|\neg H)}$$

The natural question that arises is the extent to which Good's weight satisfies the desiderata related to Keynes' notion of weight. First, let us think about weight increase with sample size. If in an experiment the observations $E_1, \ldots, E_K$ are independent given H and independent given $\neg H$, the resulting joint likelihood is the result of the multiplication of the individual likelihoods, and so the resulting joint weight is the result of adding the individual weights.

For example, suppose a die is selected at random from a hat containing nine fair dice and one loaded die with the chance 1/3 of obtaining a six. The initial uniform distribution gives you weight of evidence for the die being loaded of $log_{10}(.1)$, that is -1 (Good and Turing would say, it is -10 db). Now, every time you toss it and obtain a six, you gain $log_{10}(\frac{1/3}{1/6}) = log_{10}(2)$, that is 0.30103, and every time you toss it and obtain something else, the weight changes by $log_{10}(\frac{2/3}{5/6}) = log_{10}(.8)$, that is -0.09691. Let us inspect the weights in db (that is, multiplied by 10) for all possible outcomes of up to 20 tosses (Figure 1).

Two facts are notable. (1) Weight can drop with sample size: for instance the weight for 4 others and 5 sixes is 1.2db, and it is .2db for 5 others and 5 sixes. (2) Weight can drop while the sample size increases even if the proportion of sixes remains the same. For instance, if none of the observations are sixes, the weights go from -10 to -19.7 as the sample size goes from 0 to 10. Less trivially, the observation of one six in five leads to weight of -10.9, while the observation of two sixes in ten tosses leads to weight -11.7. That is, (Monotonicity), (Completeness), (Weak increase) and (Strong increase) all fail for Good's measure.

Moreover, there is a conceptual difficulty in the neighborhood. Suppose you are trying to ascertain the bias θ of a coin, but you do not restrict yourself to two hypotheses as in the dice example, but rather initially take any bias to be equally likely. For each particular hypothesis $\theta = x$ and any set of observations E you can use the binomial distribution to calculate $\mathsf{P}(E|\theta = x)$. But to deploy Good's definition, you also need $\mathsf{P}(E|\theta \neq x)$, which is less trivial, as now you have to integrate to calculate the expected probability of the evidence given an infinite array of possible values of y. Suppose you have no problem calculating such items. Now imagine you observe 10 heads in 20 tosses. The question 'how weighty is the evidence' makes no sense here, as Good's weight needs a hypothesis (and its negation) to be plugged in. For this reason, in such a situation, we can at best talk about a continuum of Good's weights, one for each particular value of θ.

- compare to pointwise mutual information
- evaluate in light of the desiderata

1.5 Skyrms and resilience?

1.6 Imprecision and weight with intervals

Keynes' later works and Peden's paper

[3] To be fair, logarithms of the ratio of posterior odds to prior odds have been used Jeffrey in 1936, [CITE] and the use of logarithm to ensure additivity has been suggested by Turing [CITE 1950 o 63]. Good's measure differs from Jeffrey's by taking the ratio of likelihoods rather than odds. In fact, the former ratio is identical to $O(H|E)/O(H)$, the ratio of conditional odds of H to the prior odds of H.

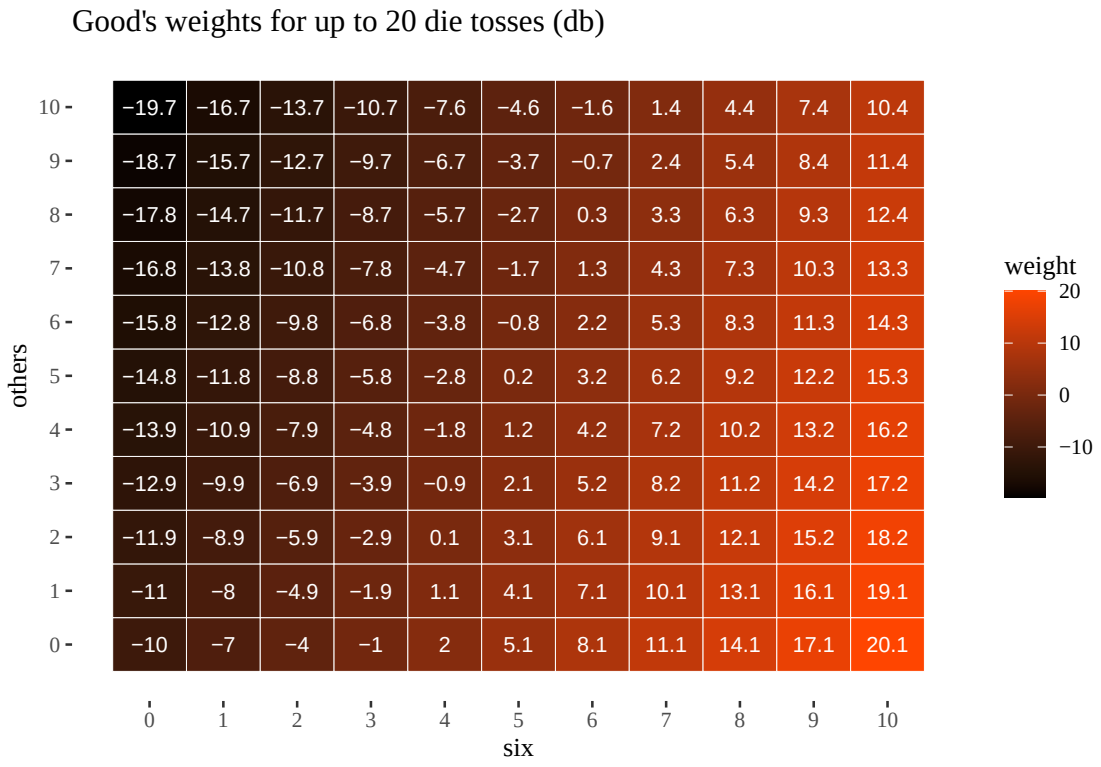

Figure 1: Good's weights in dbs, rounded, for all possible outcomes of up to 20 tosses of a die randomly selected from 10 dice nine of which were fair, and one is ⅓ loaded towards six. H ='the die is loaded'.

1.6.1 Sharpening by richness

1.6.2 Sharpening by specificity

1.6.3 Sharpening by precision

1.7 Imprecision: a second-order approach

1.8 Information-theoretic weight of evidence

1.9 Completeness tends to improve weight

1.10 Weight tends to improve accuracy

Here is a question asked by [COHEN 1986 TWELVE p. 276]: is it worth while knowing the weight of an argument without knowing its probability? In our terminology, questions inspired by Cohen's are: what's the point of weight considerations if we already have the distributions? Can weights be put to use if we do not have the distributions?

2 Literature to discuss

Kasser, 2016, Two Conceptions of Weight of Evidence in Peirce's Illustrations of the Logic of Science [DOWNLOADED]

Feduzi, 2010, On Keynes's conception of the weight of evidence [READ]

Cohen 1986, Twelve Questions about Keynes's Concept of Weight [READ]

Pedden, William 2018, Imprecise probability and the measurement of Keynes' weight of arguments

Levi 2011, the weight of argument [DOWNLOADED]

Skyrms 1977 resiliency, propensities [DOWNLOADED]

Synthese 186 (2) 2012, volume on Keynesian weight [CHECKED, NOT MUCH ON WEIGHT ACTUALLY, NO NEED TO READ]

Good, weight of evidence, survey

Good, PROBABILITY AND THE WEIGHING OF EVIDENCE

David Hamer, Probability, anti-resilience, and the weight of expectation [READ]

William Peden, Imprecise Probability and the Measurement of Keynes's "Weight of Arguments"

Runde, Keynesian Uncertainty and the weight of arguments [DOWNLOADED]

Weatherson, 2002, Keynes, uncertainty and interest rates [DOWNLOADED]

Jeffrey M. Keisler, Value of information analysis: the state of application

Edward C. F. Wilson, A Practical Guide to Value of Information Analysis

Joyce JM (2005) How probabilities reflect evidence.