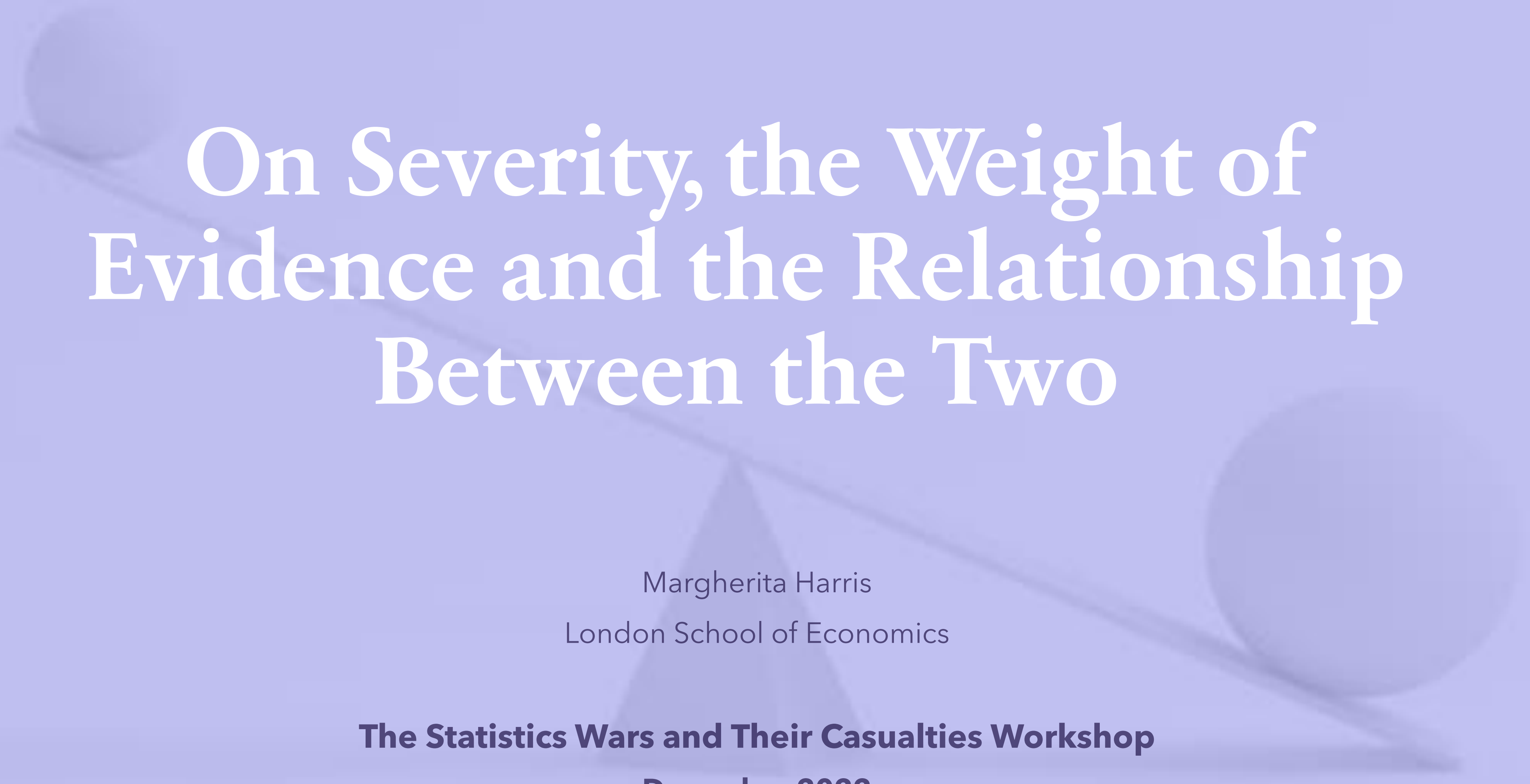


On Severity, the Weight of Evidence and the Relationship Between the Two

Margherita Harris

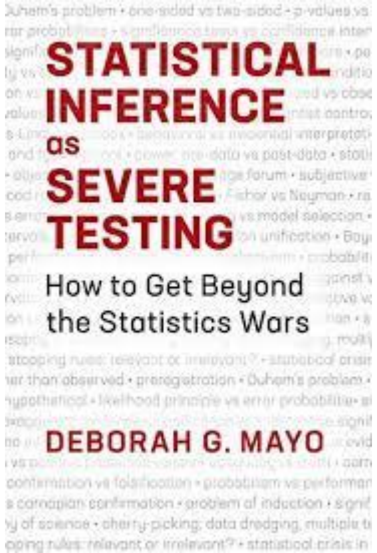
London School of Economics

The Statistics Wars and Their Casualties Workshop

December 2022

Two competing principles

Strong severity principle: We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C. (Mayo, 2018)


STATISTICAL INFERENCE as SEVERE TESTING
How to Get Beyond the Statistics Wars
DEBORAH G. MAYO

Two competing principles

Strong severity principle: We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C. (Mayo, 2018)

The likelihood principle: All the information about θ obtainable from an experiment is contained in the likelihood function for θ given x. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another. (Berger and Wolpert, 1988)

The likelihood principle: All the information about Θ obtainable from an experiment is contained in the likelihood function for Θ given x. Two likelihood functions for Θ (from the same or different experiments) contain the same information about Θ if they are proportional to one another.

What exactly does it mean to say that all of the information that data provides about a parameter Θ is contained in the likelihood function?

The likelihood principle: All the information about Θ obtainable from an experiment is contained in the likelihood function for Θ given x. Two likelihood functions for Θ (from the same or different experiments) contain the same information about Θ if they are proportional to one another.

Bayesian Confirmation Theory

Let S be a set of mutually exclusive, collectively exhaustive hypotheses, and let H be a variable ranging over members of S. Let E be a set of observed data. The likelihood function is $L(H,E)=Pr(E|H)$.

Let $c(H, E)$ indicate the incremental confirmation that E provides for H.

The likelihood principle: If the likelihood functions $L(H,E)$ and $L(H, E^*)$ are proportional (i.e. if $Pr(E|H)=kPr(E^*|H)$), then for all H in S, $c(H, E) = c(H, E^*)$.

Does Bayes' theorem entail the likelihood principle?

$$Pr(H|E) = \frac{Pr(H) \cdot Pr(E|H)}{Pr(E)}$$

No! Bayes' theorem says something about P(H|E) but nothing about c(H, E).

Does Bayes’ theorem entail the likelihood principle?

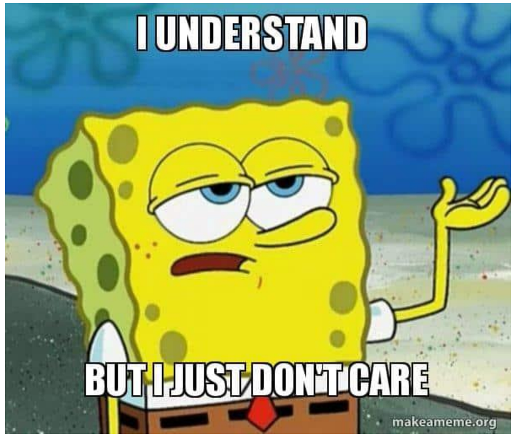
$$Pr(H|E) = \frac{Pr(H) \cdot Pr(E|H)}{Pr(E)}$$

No! Bayes' theorem says something about P(H|E) but nothing about c(H, E).

- **A consequence from Bayes’ theorem:**
  If the likelihood functions L(H, E) and L(H, E*) are proportional, then Pr(H|E) = Pr(H|E*) for every H.

- **Additional Assumption:**
  If P(H|E) = P(H|E*), then c(H, E) = c(H, E*)

⇒ **Likelihood principle**: If the likelihood functions L(H,E) and L(H, E*) are proportional, then for all H in **S**, c(H, E) = c(H, E*).



✓ **The likelihood principle**: If the likelihood functions $L(H,E)$ and $L(H, E^*)$ are proportional (i.e. if $Pr(E|H)=kPr(E^*|H)$), then for all H in S, $c(H, E) = c(H, E^*)$

✗ **Strong severity principle**: We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C.

Outline

1. Keynes's (1921) notion of the weight of evidence and the Bayesian never ending quest to account for it

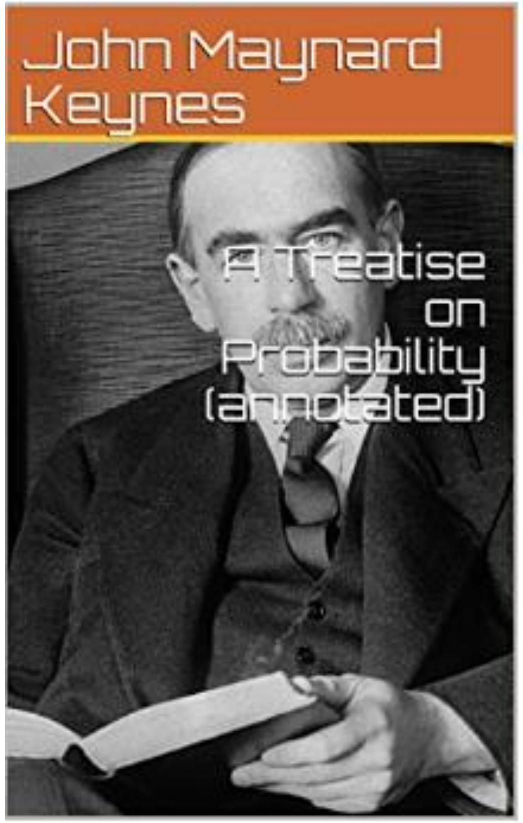
2. Claim: The likelihood principle is unsatisfactory in the Bayesian's own eyes.

3. The weight of evidence and the troubles it causes to this day:

- The literature on the burden of proof
- The conceptualization of uncertainty by the Intergovernmental Panel on Climate Change

The Weight of Evidence

"As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, 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 the argument. New evidence will sometimes decrease the probability of an argument, but it will always increase its 'weight'." (Keynes 1921, 77)

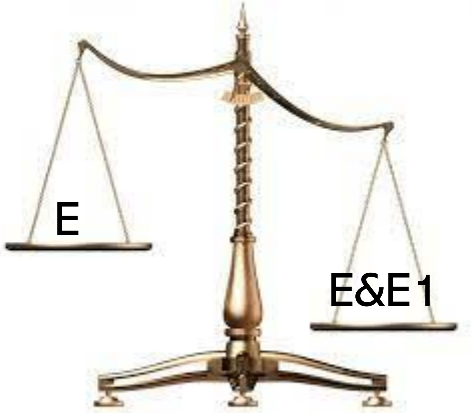


The Weight of Evidence

Can we directly measure the weight of evidence? ✕

Can we compare the weight of two distinct evidential sets?

Only under limited conditions



The Bayesian response: The weight of evidence is reflected in the resiliency of one’s credence in H

The paradox of ideal evidence (Popper,1959)

John is presented with a coin and he is asked to assign a probability to the proposition H that it will come up heads next time it is tossed. He doesn't know whether the coin is fair or whether it is biased towards heads or tails. So in light of his ignorance, he assigns a probability of 1/2 to H i.e. his subjective prior in H is $Pr(H)=1/2$. He is then allowed to toss the coin a thousand times and he gets about 50% heads and 50% tails. Call this evidence E. The probability that John assigns to the proposition H that the coin will land heads next time it is tossed is still 1/2 i.e. $Pr(H|E)=Pr(H)=1/2$.

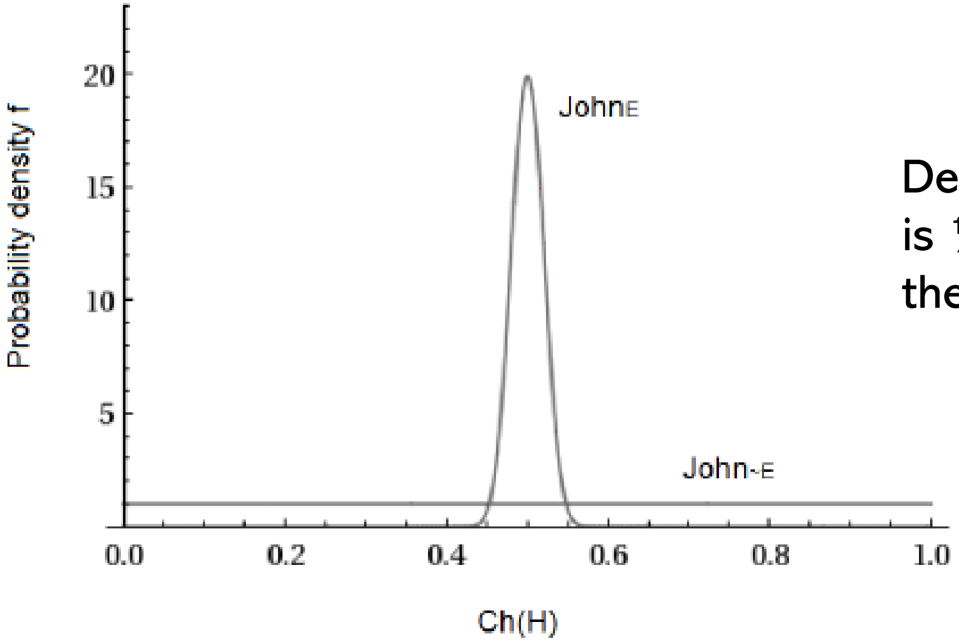
The Bayesian response: The weight of evidence is reflected in the resiliency of one's credence in H

Although John prior to observing the evidence and John after observing the evidence assign the same probability to the proposition H that the coin will come up heads the next time it is tossed, they nonetheless

> "assign different values to any proposition A(n) that asserts, concerning $n \geq 2$ distinct tosses, that all of them yield heads. To any such proposition [John after having observed E] assigns the value $\left(\frac{1}{2}\right)^n$; but to the same proposition [John prior to having observed E] must assign a higher value, if you hope to learn from experience" (Jeffrey 1965, 196).

"The ideal evidence has changed not the probability of [heads] on toss a, but rather the resiliency of the probability of [heads] on toss a." Skyrms (1977, 707)

The Bayesian response: The weight of evidence is reflected in the resiliency of one's credence in H



Despite this difference, however, both John's credence in H is ½ in both cases since John's expectation of the chance of the coin landing heads is 1/2 in both cases

$$Pr(H) = Pr(H|E) = \int_0^1 f(Ch(H) = x) \cdot x\,dx = 1/2$$

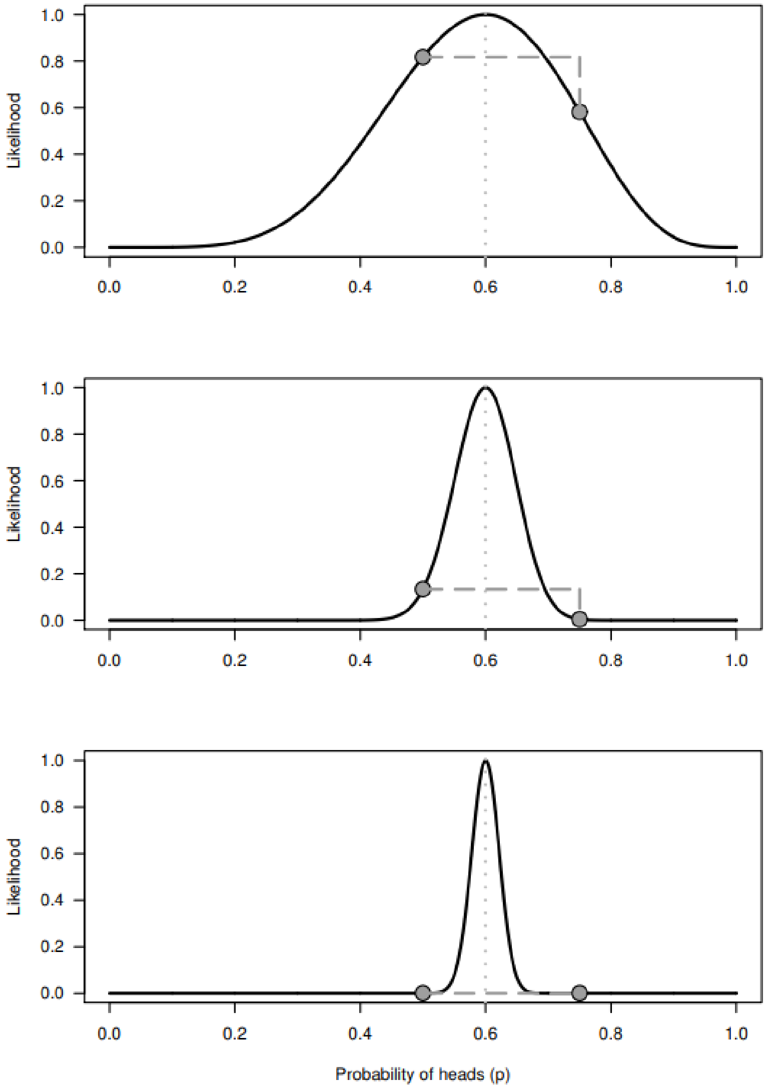
“This situation provides no analogy for a growing body of heterogeneous evidence in juridical proof… Suppose in a wrongful death case the court was seeking to determine whether, without the defendant’s negligence, the deceased would have continued working until 65 years. [In this case] there appears no basis to expect the evidence or the probability assessment it generates to conform to a regular pattern. The deceased smoked a packet of cigarettes a day, suggesting low life expectancy. But he exercised regularly, suggesting a high life expectancy. But he worked as an industrial chemist with the risk of exposure to highly toxic substances. But his parents are still living and in good health and his grandparents all lived to 90 years. But he raced motorbikes. But his diet was extremely good. And so on. …in this longevity hypothetical there is no apparent reason to expect the probability assessment to converge towards a particular point on the unit interval as the weight of evidence increases.” (Hammer, 2012)

The likelihood function of p

6 heads in 10 flips

60 heads in 100 flips

300 heads in 500 flips



If the Bayesian thinks, as they seem to think, that there is something to the notion of the weight of evidence, and yet the weight of evidence is not always reflected in the likelihood function of a hypothesis, then this should be enough of a reason for a Bayesian to reject the likelihood principle.

If the Bayesian thinks, as they seem to think, that there is something to the notion of the weight of evidence, and yet the weight of evidence is not always reflected in the likelihood function of a hypothesis, then this should be enough of a reason for a Bayesian to reject the likelihood principle.

The Weight of evidence and Severity

The weight will increase as the data increases, regardless of how that data was acquired. But to the severe tester what matters in the assessment of whether one has evidence in support of a hypothesis is not how much relevant evidence one has gathered per se, rather it is whether or not the process of generating that evidence has been able to severely test the hypothesis in question.

If the Bayesian thinks, as they seem to think, that there is something to the notion of the weight of evidence, and yet the weight of evidence is not always reflected in the likelihood function of a hypothesis, then this should be enough of a reason for a Bayesian to reject the likelihood principle.

The Weight of evidence and Severity

The weight will increase as the data increases, regardless of how that data was acquired. But to the severe tester what matters in the assessment of whether one has evidence in support of a hypothesis is not how much relevant evidence one has gathered per se, rather it is whether or not the process of generating that evidence has been able to severely test the hypothesis in question.

A mere symptom…

The Bayesian's recognition that the posterior probability that one assigns to a hypothesis conditional on the available evidence is unable to reflect something that to the Bayesian herself seems important about the nature of that evidence (i.e. its weight) is merely a symptom of the Bayesian's *own* dissatisfaction with the likelihood principle.

The weight of evidence and the burden of proof

A general discomfort with the idea that all that is required for a guilty verdict is that the available evidence makes it highly probable that the accused is guilty:

“ [the probabilistic measure] may say something about the evidence to hand, but it says nothing about the completeness or weight of that evidence. A slight body of evidence may give rise to a reasonable degree of probabilistic certainty, but what of the fragility of that assessment, given the many items of further evidence that have not yet been considered?” (Hamer, 2012).

Davidson and Pargetter (1987):

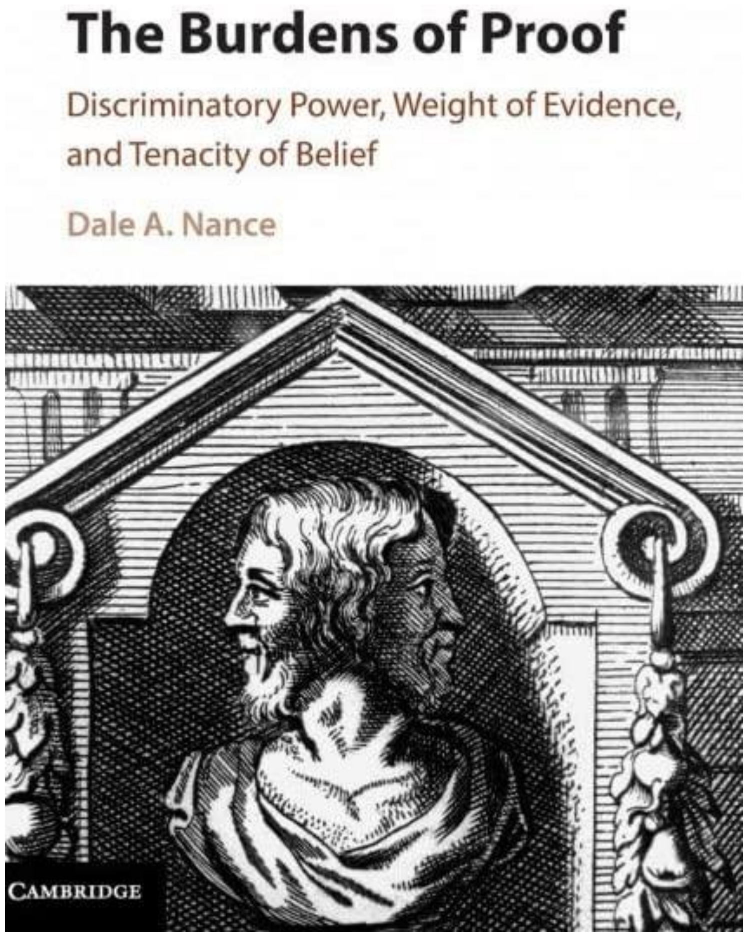
“Cases can be dismissed on the grounds of insufficient evidence, or the guilty verdict resisted if there is insufficient evidence, even if the evidence that is available is reliable and does make it highly probable that the accused is guilty. This suggests that guilt beyond reasonable doubt is only established if the jury believes that they have all the relevant evidence, in the sense that any further evidence which probably obtains would not change the probability of guilt; or alternatively, that any evidence which would lower the probability of guilt has a low probability.

Our third requirement for guilt beyond reasonable doubt is thus

(c) the probability of guilt must have high resilience relative to all bodies of evidence which are probable.

We shall call strength of evidence in this sense ‘weight', and so our third requirement is that the probability of guilty should have high weight. We follow J. M. Keynes use of the term 'weight' to refer to this feature of a probability”

"It is a mistake, [….] to *identify* the idea of Keynesian weight with the idea of resilience…. The two ideas are distinct, and though there are causal connections between them, the dependence relationships are complex. We should not, therefore, embrace the suggestion by Davidson and Pargetter that Keynesian weight be *identified with* resilience of probability […]"



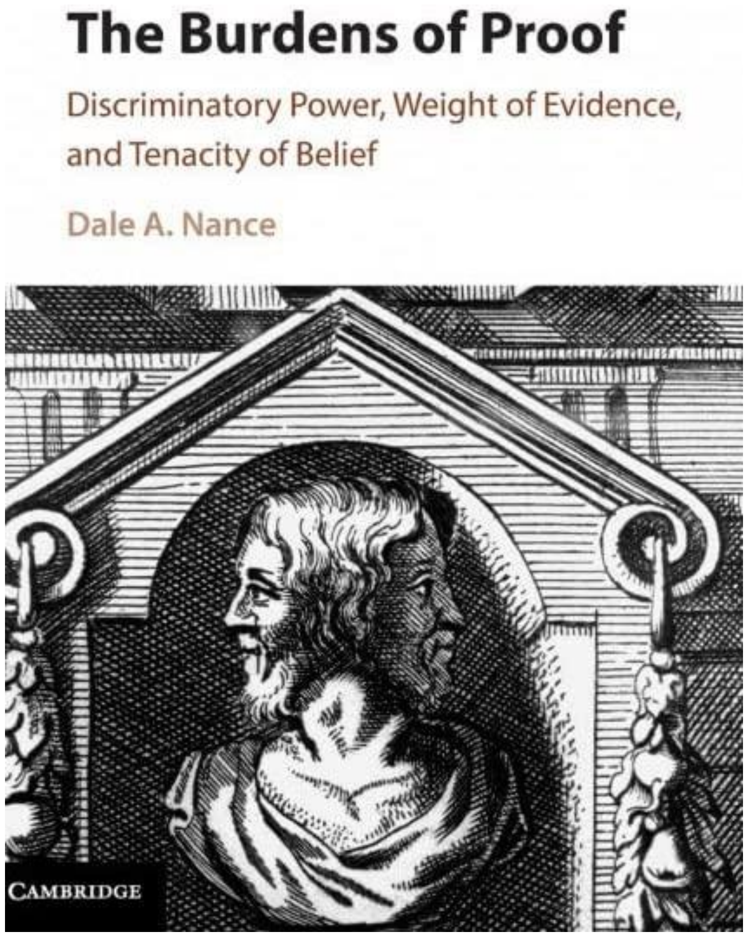
The Burdens of Proof

Discriminatory Power, Weight of Evidence, and Tenacity of Belief

Dale A. Nance

CAMBRIDGE

“Despite all this, the concept of resilience does have some use… the harder it is to shift the probability of interest, that is, the more resilient that probability is, the less is the expected gain in utility of determining which of various possible states of an evidenced fact obtains before acting, and so the more rational it is to act on the probability one has at the time. Resilience, therefore, is an ingredient in the rationality of acting on a probability by being an ingredient in deciding whether one should augment Keynesian weight by getting more information before acting on one’s assessed probability. This connection places resilience in its proper place; in the legal context, it is a component part of the decision by the court whether the case is ripe for submission of the dispute for decision by the fact-finder…. Other factors include the cost of the acquisition of further information […] the extent to which the probability falls below or exceeds some threshold of decision [….]”



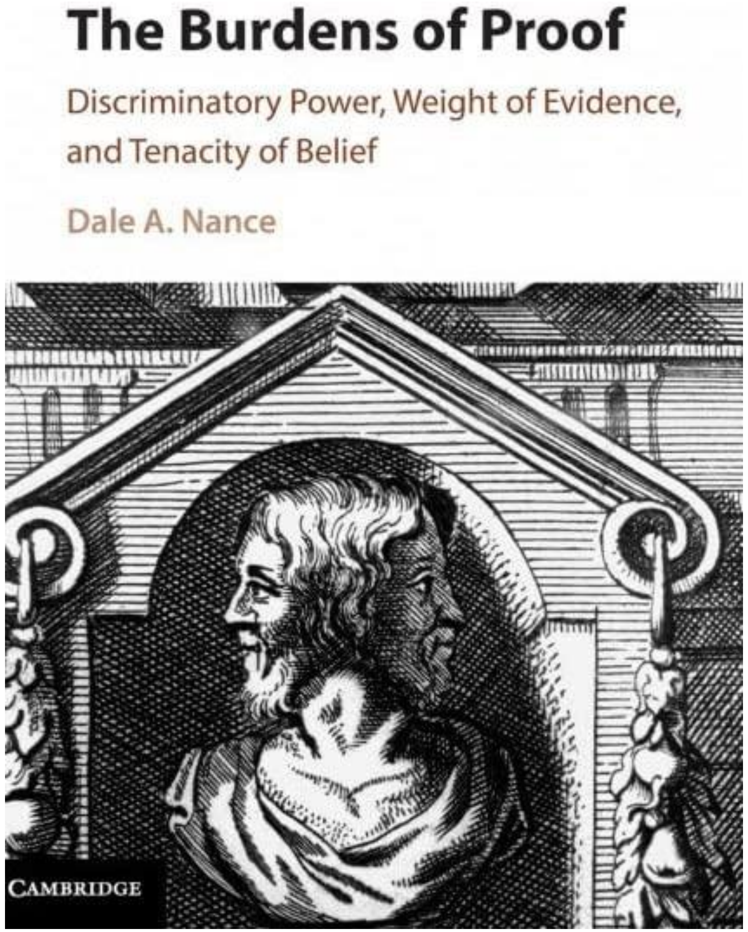
The Burdens of Proof

Discriminatory Power, Weight of Evidence, and Tenacity of Belief

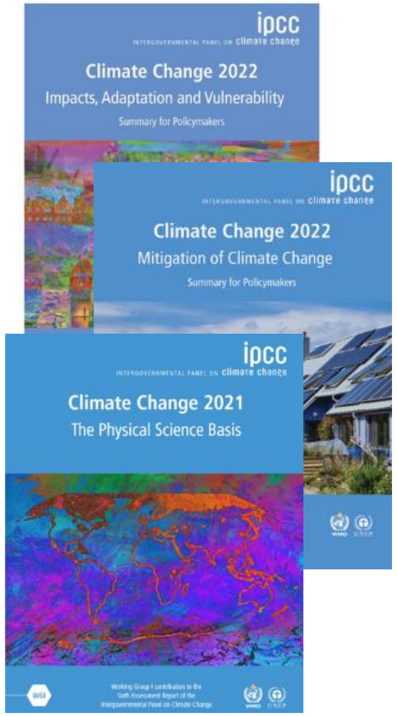
Dale A. Nance

CAMBRIDGE

Dale's attempt to account for the weight of evidence, together with all other attempts, are bound to be unsatisfactory from an epistemic point of view. In Dale's case, his requirement of practical optimisation of Keynesian weight is clearly epistemically unsatisfactory, because it suggests that whether or not one is justified in declaring to have sufficient evidence for a claim depends on *pragmatic* factors that have nothing to do with the nature of the evidence at one's disposal nor with how severely that claim has been put to test.


The Burdens of Proof
Discriminatory Power, Weight of Evidence, and Tenacity of Belief
Dale A. Nance
CAMBRIDGE

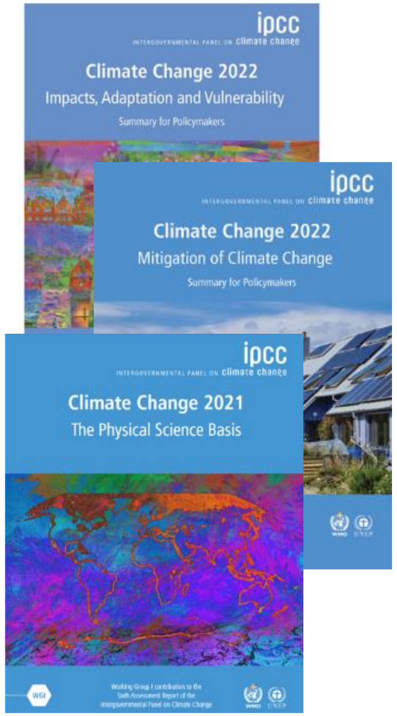
Intergovernmental Panel on Climate Change (IPCC)


ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2022
Impacts, Adaptation and Vulnerability
Summary for Policymakers
ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2022
Mitigation of Climate Change
Summary for Policymakers
ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2021
The Physical Science Basis
WGI
Working Group I contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change
WMO
UNEP

“Equilibrium climate sensitivity is *likely* in the range 1.5°C to 4.5°C (*high confidence*).

“Relative to the average from year 1850 to 1900, global surface temperature change by the end of the 21st century [. . . is] *unlikely* to exceed 2°C for RCP2.63 (*medium confidence*).”

Intergovernmental Panel on Climate Change (IPCC)


ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2022
Impacts, Adaptation and Vulnerability
Summary for Policymakers
ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2022
Mitigation of Climate Change
Summary for Policymakers
ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2021
The Physical Science Basis

“Equilibrium climate sensitivity is *likely* in the range 1.5°C to 4.5°C (*high confidence*).

“Relative to the average from year 1850 to 1900, global surface temperature change by the end of the 21st century [. . . is] *unlikely* to exceed 2°C for RCP2.63 (*medium confidence*).”

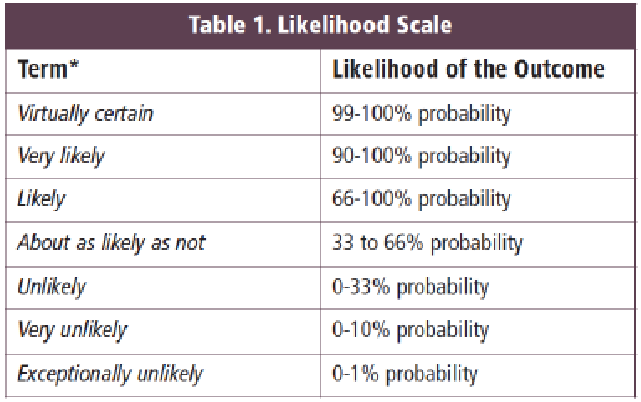
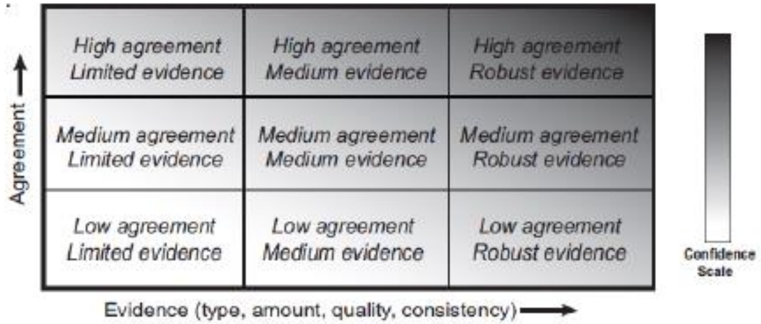


Table 1. Likelihood Scale

Term*	Likelihood of the Outcome
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

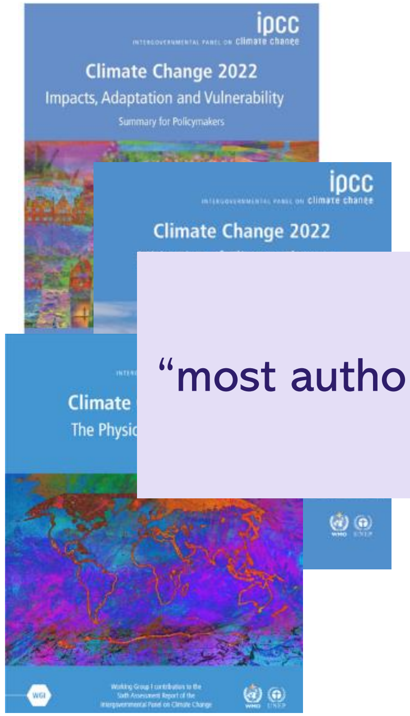
FIGURE 1.2: The likelihood metric



High agreement *Limited evidence*	*High agreement* *Medium evidence*	*High agreement* *Robust evidence*
Medium agreement *Limited evidence*	*Medium agreement* *Medium evidence*	*Medium agreement* *Robust evidence*
Low agreement *Limited evidence*	*Low agreement* *Medium evidence*	*Low agreement* *Robust evidence*

FIGURE 1.1: ‘A depiction of evidence and agreement statements and their relationship to confidence. Confidence increases towards the top-right corner as suggested by the increasing strength of shading.’ (ibid., 3)

Intergovernmental Panel on Climate Change (IPCC)


ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2022
Impacts, Adaptation and Vulnerability
Summary for Policymakers
ipcc
INTERGOVERNMENTAL PANEL ON climate change
Climate Change 2022
Climate
The Physic

“Equilibrium climate sensitivity is *likely* in the range 1.5°C to 4.5°C (*high confidence*).

“Relative to the average from year 1850 to 1900, global surface temperature change by the end of the 21st century [. . . is] *unlikely* to exceed 2°C for RCP2.63 (*medium confidence*).”

“most authors agreed that the confidence/likelihood distinction was confusing” (Janzhood, 2020).

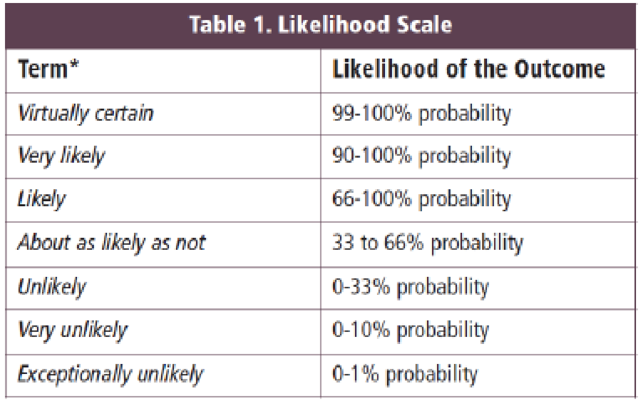
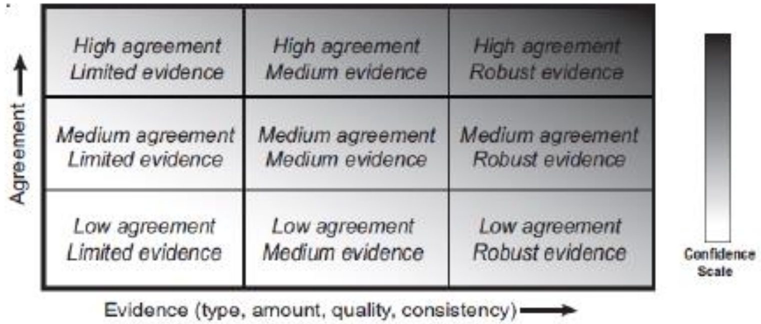


Table 1. Likelihood Scale

Term*	Likelihood of the Outcome
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

FIGURE 1.2: The likelihood metric

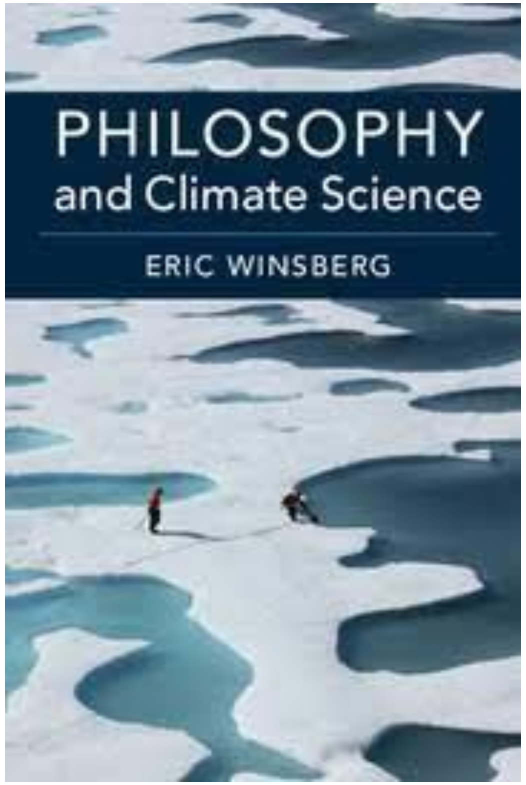


High agreement *Limited evidence*	*High agreement* *Medium evidence*	*High agreement* *Robust evidence*
Medium agreement *Limited evidence*	*Medium agreement* *Medium evidence*	*Medium agreement* *Robust evidence*
Low agreement *Limited evidence*	*Low agreement* *Medium evidence*	*Low agreement* *Robust evidence*

FIGURE 1.1: ‘A depiction of evidence and agreement statements and their relationship to confidence. Confidence increases towards the top-right corner as suggested by the increasing strength of shading.’ (ibid., 3)

"Equilibrium climate sensitivity is *likely* in the range 1.5°C to 4.5°C (*high confidence*)"

"One possible way to understand measures of confidence might be 'as a kind of second-order probability'; that is to say, the high confidence in the imprecise credence [0.66, 1] above is a bit like a high degree of belief that the credence will be resilient in the face of future evidence- assessed by looking at the variety of evidence supporting the credence, and the degree of agreement among those sources supporting the credence." (Winsberg, 2018)



Some casualties and a cry for help...