# Vassend on Verisimilitude and Counterfactual Probabilities

Niiniluoto, Ilkka

2021-07-01

Niiniluoto , I 2021 , ' Vassend on Verisimilitude and Counterfactual Probabilities ' , Philosophy of Science , vol. 88 , no. 3 , pp. 554-561 . https://doi.org/10.1086/712883

http://hdl.handle.net/10138/345754
https://doi.org/10.1086/712883

# Vassend on Verisimilitude and Counterfactual Probabilities

Ilkka Niiniluoto*

Olav Benjamin Vassend proposes two solutions to the "interpretive problem" of assigning nonzero probabilities to hypotheses that are known to be false. He argues that the verisimilitude interpretation (probability expresses the degree of belief that the hypothesis is closest to the truth) and the counterfactual interpretation (probability is conditional on a false supposition) are equivalent. While Vassend's intuition about these two solutions is basically correct, the technical details of his treatment need elaboration and correction. Appropriate tools for combining verisimilitude and Bayesian probabilities can be found in my *Truthlikeness*.

**1. Vassend on the Interpretive Problem.** Vassend (2019) formulates the *interpretive problem* by noting that Bayesian methods often assign nonzero probabilities to hypotheses that are known to be false. This is inconsistent with the standard interpretation that Bayesian probabilities express degrees of belief in the *truth* of a hypothesis given evidence. Vassend argues that this problem can be solved only by giving a new semantics for Bayesian inference.

Let $\Theta$ be a set of rival and mutually exclusive hypotheses indexed by a parameter $\theta$, and let X be the set of possible outcomes of an observation or experiment. Then a Bayesian statistical model specifies the prior probability $p(\theta)$ of $\theta$ in $\Theta$ and the likelihood $p(x/\theta)$ of $\theta$ on x in X, and calculates the posterior probability $p(\theta/x)$ of $\theta$ given x by means of Bayes's theorem as proportional to the product $p(\theta)p(x/\theta)$. Assume now that background information K implies that each hypothesis $\theta$ in $\Theta$ is false. Then, given K, $p(\theta)$ as the degree of belief in the truth of $\theta$ is zero, which implies by Bayes's theorem that $p(\theta/x)$ is zero as well. In this situation Bayesian inference collapses.

Received November 2019; revised March 2020.

*To contact the author, please write to: Department of Philosophy, History, and Art Studies, 00014 University of Helsinki, Finland; e-mail: ilkka.niiniluoto@helsinki.fi.

Philosophy of Science, 88 (July 2021) pp. 554–561. 0031-8248/2021/8803-0009$10.00

Vassend's *verisimilitude interpretation* suggests that the prior p($\theta$) is understood as

the probability that $\theta$ is closest to the truth out of the hypotheses in $\Theta$. (1)

If a verisimilitude measure v is available, then according to (1) p($\theta$) is the probability that $\theta$ maximizes v. Even though typically v is maximized by a single hypothesis in $\Theta$, so that disjunctive hypotheses do not satisfy the condition (1), the probability of members of $\Theta$ can be formally extended to their exclusive disjunctions by the addition law.

The *counterfactual interpretation*, which Vassend attributes to Sprenger's (2017) unpublished paper, states that the prior probability of $\theta$ should be understood as the conditional probability p($\theta$/b), where b is the counterfactual assumption that one of the hypotheses in $\Theta$ is true. In statistical regression analysis, b might be the false idealizing assumption that the functional dependency of two quantities can be expressed by a linear function.

While Sprenger seems to think that these two alternatives are separate interpretations of probabilities, Vassend's contribution is to argue that his verisimilitude interpretation is in fact "intertranslatable" with the counterfactual interpretation. The reason why these approaches are "two sides of the same coin" comes primarily from the fact that the semantics of counterfactuals involve a similarity ranking s of possible worlds, and the functions v and s can be defined in terms of each other.

Vassend's project has an aim similar to Shimony's (1970) "tempered personalism," which wishes to assign nonzero "degrees of rational commitment" to "seriously proposed hypotheses," even if their degree of belief in the ordinary sense is zero or extremely small. Shimony too typically has in mind situations in which the true theory is not included among the relevant hypotheses, but his treatment is clearly different from Vassend's.

**2. The Problem with Sharp Hypotheses.** Let us start the assessment of Vassend's article by noting a special issue with *sharp* hypotheses. These are typically statements that pick out a point in a continuum, so that their probability measure is zero. Vassend thinks that his set of hypotheses $\Theta$ can be assumed to be finite "without loss of generality" (2019, 699), but this is potentially misleading, since, for example, in the estimation of real-valued parameters $\Theta$ is a continuous subset of the infinite class R of real numbers. Then the prior probability density p($\theta$) may have a nonzero value for each point $\theta$ in $\Theta$, but the probability of a sharp hypothesis $\theta$ is nevertheless zero. This situation is different from Vassend's interpretive problem, since all sharp hypotheses in $\Theta$ receive zero probabilities—independently of the question whether $\Theta$ includes a true element. But the interpretive problem is in a sense independent of the problem with a sharp hypothesis, since it reappears as the task of making sense of nonzero probability densities for point hypotheses and nonzero probabilities for regional hypotheses, when all the hypotheses are known to be false.

When Vassend illustrates the interpretive problem in practice in his section 3, he chooses the class $\Theta$ so that it includes all functions of the form $Y = \alpha(1{,}010 - X)^n + \varepsilon$ between quantities X and Y. In section 6 he simplifies this equation to straight lines. In both cases $\Theta$ is a continuous subset of the real space (three parameters in $R^3$ or two parameters in $R^2$), so that the counterfactual probability of its individual members is zero.

Even though sharp hypotheses have zero probability, their degree of *probable approximate truth* may be nonzero (see Niiniluoto 1987, 280). This degree can be defined for a real-valued hypothesis $\theta$ as the probability of a small interval $u(\theta)$ around $\theta$, that is, the sum of all values $p(\theta)$ for $\theta$ in $u(\theta)$. The same idea can be generalized to hypotheses that are real-valued curves. This concept, which has important applications in Bayesian probably approximately correct learning as a method in machine learning (see Niiniluoto 2005), is clearly different from Vassend's verisimilitude interpretation of probability.

**3. Verisimilitude Measures.** Standard expositions of the similarity approach to verisimilitude define this concept relative to a cognitive problem B, which consists of a partition of mutually exclusive and jointly exhaustive hypotheses $h_i$ in some interpreted language L (see Oddie 1986; Niiniluoto 1987). Then there is one and only one element $h^*$ in B that is true (in the actual world). This $h^*$ is the unknown element of the problem B, and the task of the investigator is to identify it. If a metric or distance function d is defined between the elements of B, and d is extended to measure the distances of disjunctions of hypotheses $h_i$ to an element in B, then a *verisimilitude measure* v is obtained for any theory H expressible in L: the degree of truthlikeness $v(H, h^*)$ of H depends on its distance to the target $h^*$.

The choice of the distance function d depends on the type of hypotheses in the cognitive problem B. Vassend's section 6 illustrates this in the case in which the compared hypotheses are real-valued curves in $R^2$, but his attempt is not successful. Niiniluoto (1987, 385–86) proposes to measure the distance between two real-valued quantitative laws f and g by the Minkowski metrics for function spaces:

$$L^p(f, g) = \left[ \int | f(z) - g(z)|^p dz \right]^{1/p}, \tag{2}$$

where integration is defined over the domain of the functions f and g. Here $L^p$ for small values of p reflects the area between the curves f and g, while for infinite p it gives the maximum distance between the values of f and g. Apart from this maximum distance, Vassend mentions the minimum distance between f and g (2019, 708), but for a good reason this is not included in the Minkowski family (2). Indeed, the minimum distance as a measure of verisimilitude is here worthless: if any curve that intersects the true curve g has maximal verisimilitude, then we have no useful discrimination or ordering among false hypotheses by this criterion. As a curve is a conjunction of an infinite number of claims, also the maximum distance is a crude measure, and the Manhattan metric ($p = 1$) and the Euclidean metric ($p = 2$) are more plausible.

**4. Verisimilitude with a False Presupposition.** According to Vassend's verisimilitude interpretation (1), $p(\theta)$ is the probability that $\theta$ is closest to the truth among hypotheses in $\Theta$, when background knowledge implies the falsity of all elements of $\Theta$. But this means that $\Theta$ is not a cognitive problem in the standard sense, since there is no element in $\Theta$ that serves as the true target of the verisimilitude measure v. Thus, the condition (1) that $p(\theta)$ is the probability that $\theta$ maximizes the verisimilitude measure v is not well defined.

A solution to this problem has been proposed by Niiniluoto (1987, 259–62). Let $B_b = \{h_i | i \varepsilon I\}$ be a cognitive problem in language L relative to a false presupposition b. Thus, b implies that the elements $h_i$ in $B_b$ are mutually exclusive and jointly exhaustive, but they all are false. If s is a metric on the space of L-structures (possible worlds), then choose the target $h^*[b]$ as the element of $B_b$ that is true in the b-world that is minimally distant from the actual world. If there is more than one such minimally distant L-structure where b is true, choose the target as the disjunction of the corresponding statements in $B_b$.[1] Using the semantics of counterfactuals by Lewis (1973), this choice can be expressed by saying that the target $h^*[b]$ of a cognitive problem $B_b$ with a false presupposition b is the most informative statement in the language of $B_b$ that *would be true if b were true*. By this choice, the elements of $B_b$ and their disjunctions can be compared for verisimilitude $v_b$ by their

1. A referee of this journal finds the proposed "proxy solution" problematic. If we are looking for the geographic center of Italy and Florence happens to be slightly closer to this center than Rome, then we would measure the closeness to the center by the distance to Florence, evidently disadvantaging southern Italian places. It is not clear what the counterfactual assumption is in this example, but if it is the restriction of choices to big cities, then this idealization is not fruitful. It should be acknowledged that idealizations exclude exact truth from our set of available hypotheses and thereby are biased in one way or other. But, as Galileo was well aware, idealizations may be methodologically useful, if their concretizations or de-idealizations help in the search for truth. So here the idealized network of big cities could be replaced by a rectangular grid, whose system of nodes can be made tighter step by step, so that eventually we are approximating the true center of Italy.

distance to the target in $B_b$. And the rule (1) now states that the prior $p(h_i)$ of $h_i$ in $B_b$ expresses the probability that $h_i$ is identical with this target $h^*[b]$.

Another solution is implicit in Vassend's paper.[2] Assume that $B_b$ is a subclass of a larger cognitive problem B with the true target $h^*$, where B is defined without the counterfactual assumption b. Then the members of $B_b$ can be compared for their verisimilitude by the measure v for B (instead of $v_b$ for $B_b$), even though the target $h^*$ is not included in the set $B_b$. There may be issues about the choice of B, and the comparison of verisimilitude orderings v and $v_b$, but for the purposes of Vassend's project they are irrelevant, since both v and $v_b$ are maximized in $B_b$ by the hypothesis $h^*[b]$. According to both solutions, the probability that an element $h_i$ of $B_b$ is closest to the truth is the probability that $h_i$ is $h^*[b]$. Note also that if b is true, then $h^*[b]$ is identical with $h^*$.[3]

Vassend correctly notes that similarity rankings of possible worlds and verisimilitude rankings of propositions induce each other (2019, 712). It is also correct that this sort of link was used by Hilpinen (1976) in his pioneering account of the similarity approach to truthlikeness (n. 15). But Vassend's exposition here is flawed. Hilpinen employed Lewis's qualitative notion of similarity spheres as a primitive, while Niiniluoto (1987) explicitly defines a quantitative measure by structural similarity considerations.[4] Moreover, if theory H is a set (disjunction) of possible worlds, then the verisimilitude of H should not be defined by the minimum distance of H from the actual world, as Vassend (2019, 712) assumes.[5] Hilpinen defined the notion of *approximate truth* by the minimum distance, but as a definition of verisimilitude this is known to be hopeless. Indeed, Hilpinen argued convincingly that the comparative notion of *truthlikeness* should take into account at least the maximum distance as well, but this gives only a partial ranking of hypotheses. A quantitative version of Hilpinen's proposal was given by Niiniluoto's (1977) min-max measure, and more elaborate alternatives like Oddie's (1986) average measure and Niiniluoto's (1987) min-sum measure are still debated (e.g., Oddie 2014; Niiniluoto 2020).

2. I am grateful to Vassend (pers. comm., February 14, 2020) for the clarification of his position.

3. Note also that we are not replacing the hypothesis $h_i$ with $\langle h_i$ is closest to the truth$\rangle$ (e.g., Vassend 2019, 704).

4. This also means that possible worlds as full-blown metaphysical entities are replaced by more accessible logical tools, like Jaakko Hintikka's "small worlds" or "constituents" (maximally informative descriptions of possible worlds in a given linguistic framework). See Niiniluoto (1987), 204–9.

5. This assumption is explicit in Vassend's construction of v from s. The other direction from v to s is ambiguous, since different measures of verisimilitude might disagree on the question of which hypothesis H containing a word w is most truth-like.

**5. Counterfactual Probabilities and Expected Verisimilitude.** The preceding section gives support to Vassend's contention that probabilities (1) under the verisimilitude interpretation are equivalent to counterfactual probabilities, when the set of hypotheses $\Theta$ is based on a false presupposition. It is also correct that thereby Bayesian probabilities depend on the pragmatic factor related to the choice of the verisimilitude measure, since different distance measures s and s′ could lead to different targets $h^*[b]$ in $B_b$.

But it should be remembered that after all the Bayesian agent does not know which hypothesis is the true one (in the standard case) or would be true (in the counterfactual case), so that she may choose the priors freely as long as probability axioms are satisfied. In the standard special case in which $\Theta$ includes a true target $h^*$, a hypothesis $\theta$ has the minimum distance from the truth if and only if $\theta$ is identical with $h^*$. In the counterfactual case, the best hypothesis is identical with the target $h^*[b]$. Condition (1) is not an operational rule for finding the maximally truth-like hypothesis, but it suggests that prior probabilities should express the agent's best guesses about the location of the unknown target $h^*$ or $h^*[b]$ within the relevant cognitive problem.

Formally, let H be a theory in a cognitive problem B, and let $\varepsilon > 0$ be a small real number. Let $U_\varepsilon(H)$ be the set of hypotheses $h_i$ in B such that $v(H, h_i) \geq 1 - \varepsilon$. Then one may define *probable verisimilitude* in the following way: the probability $pv_{1-\varepsilon}(H/e)$ that the degree of truthlikeness of H is at least $1 - \varepsilon$ equals the probability of $U_\varepsilon(H)$ (see Niiniluoto 1987, 279). Then $pv_{1-\varepsilon}(H/e)$ has the maximum value 1 if and only if the posterior probability is wholly concentrated on $U_\varepsilon(H)$. It is thus possible that more than one hypothesis $h_i$ in B receives the maximum value, and it may happen that $p(h_i/e) = 0$ but $pv_{1-\varepsilon}(h_i/e) = 1$. But when $\varepsilon \rightarrow 0$, we have in the limit $pv_{1-\varepsilon}(h_i/e) \rightarrow p(h_i/e)$ for all i.

There is a further concept that behaves formally in a way different from probability, that is, *expected verisimilitude* ver(H/e), introduced by Niiniluoto (1977):

$$\mathrm{ver}(H/e) = \sum_{i \varepsilon I} p(h_i/e) v(H, h_i), \qquad (3)$$

where e is the available evidence. Here $v(H, h_i)$ is the degree of truthlikeness H would have if $h_i$ were true, which is weighted by the posterior probability $p(h_i/e)$ of $h_i$ given e, and the sum goes over all hypotheses $h_i$ in B. If $\Theta$ is a continuous space, then in (3) p is a probability density, and the finite sum operator $\Sigma$ is replaced by the integral

$$\mathrm{ver}(H/x) = \int_\Theta p(\theta/x) v(H, \theta)\, d\theta. \qquad (4)$$

(See Niiniluoto 1987, 269). Definitions (3) and (4) are tools for estimating unknown degrees of truthlikeness, when we have an epistemic probability distribution over the relevant alternatives $h_i$ in B. Unlike posterior probability P(H/e), the value of ver(H/e) may be nonzero and even high, when e contradicts H (274). If posterior probability is almost completely allocated to a particular hypothesis $h_j$ in B, so that $h_j$ is believed to be true, then the expected verisimilitude of other hypotheses $h_i$ are approximately equal with their closeness to this presumed truth $h_j$ (273).

If a cognitive problem $B_b$ is defined relative to a counterfactual presupposition b, so that the realist condition ¬b implies the falsity of all of its elements $h_i$, then $p(h_i/e\&\neg b) = 0$ and $ver(h_i/e\&\neg b) = 0$ for any true evidence e. For example, $h_i$ might describe the ballistic behavior of projectiles near the surface of the air, where b is the idealizing assumption that the resistance of air has no influence, and e describes observations under real conditions on air. Niiniluoto (1987, 286) argues that still we may have $p(h_i/e'\&b) > 0$, where e′ tells what the data e would have been in the idealized situation b. Thus, e′ is obtained—perhaps using theoretical background assumptions—by subtracting the influence of air from the observational data e (e.g., Suppes 1962) or by trying to realize the idealized circumstances in controlled experiments or computer simulations (e.g., study of free fall in a vacuum). For empty evidence, this positive prior probability $p(h_i/b)$ is an instance of Sprenger's (2017) counterfactual interpretation of conditional probability. Applying (3) with counterfactual probabilities, we may have $ver(h_i/e'\&b) > 0$ and likewise $ver(H/e'\&b) > 0$ for a disjunctive theory H in $B_b$. An alternative approach is to transform the idealized hypotheses into a realist one $h_i'$ by concretization, that is, by adding the influence of the air to the description $h_i$ (see Niiniluoto 1987, 116–17). Then we may have $p(h_i'/e\&\neg b) > 0$ and $ver(h_i'/e\&\neg b) > 0$. In other words, idealized hypotheses may be confirmed by idealized evidence, and factual hypotheses by factual evidence.

These considerations illustrate the important feature of Bayesian modeling that counterfactual assumptions may have a significant role in the determination of both prior and posterior probabilities—as well as evidence-based estimates of verisimilitude.

## REFERENCES

Hilpinen, Risto. 1976. "Approximate Truth and Truthlikeness." In *Formal Methods in the Methodology of the Empirical Sciences*, ed. Marian Przelecki, Klemens Szaniawski, and Ryszard Wojcicki, 19–42. Dordrecht: Reidel.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

Niiniluoto, Ilkka. 1977. "On the Truthlikeness of Generalizations." In *Basic Problems in Methodology and Linguistics*, ed. R. E. Butts and Jaakko Hintikka, 121–47. Dordrecht: Reidel.

———. 1987. *Truthlikeness*. Dordrecht: Reidel.

———. 2005. "Inductive Logic, Verisimilitude, and Machine Learning." In *Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress*, ed. Petr Hajek, Luis Valdés-Villanueva, and Dag Westerståhl, 295–314. London: King's College.

———. 2020. "Truthlikeness: Old and New Debates." *Synthese* 197:1581–99.

Oddie, Graham. 1986. *Likeness to Truth*. Dordrecht: Reidel.
———. 2014. “Truthlikeness.” In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. https://plato.stanford.edu/entries/truthlikeness/.
Shimony, Abner. 1970. “Scientific Inference.” In *The Nature and Function of Scientific Theories*, ed. R. C. Colodny, 79–172. Pittsburgh: University of Pittsburgh Press.
Sprenger, Jan. 2017. “Conditional Degrees of Belief.” Unpublished manuscript, PhilSci Archive. http://philsci-archive.pitt.edu/13515/.
Suppes, Patrick. 1962. “Models of Data.” In *Logic, Methodology and Philosophy Science: Proceedings of the 1960 International Congress*, ed. E. Nagel, P. Suppes, and A. Tarski, 252–61. Stanford, CA: Stanford University Press.
Vassend, Olav Benjamin. 2019. “New Semantics for Bayesian Inference.” *Philosophy of Science* 86 (4): 696–718.