Two Impossibility Results for Measures of Corroboration

January 30, 2015

Contents

Abstract

According to influential accounts of scientific method, e.g., critical rationalism, scientific knowledge grows by repeatedly testing our best hypotheses. But despite the popularity of hypothesis tests in statistical inference and science in general, their philosophical foundations remain shaky. In particular, the interpretation of non-significant results—those that do not refute the tested hypothesis—poses a major philosophical challenge. To what extent do they corroborate the tested hypothesis or provide a reason to accept it?

Karl R. Popper sought for measures of corroboration that could adequately answer this question. According to Popper, corroboration is different from probability-raising, and grounded in the predictive success and testability of a hypothesis. As such, corroboration becomes an indicator of the scientific value of a hypothesis and guides our practical preferences over hypotheses which have been subjected to severe tests.

This paper proves two impossibility results for corroboration measures that are specified along the above lines. The generality of these results shows that Popper's qualitative characterization of corroboration must be misguided. I explore what a more promising, and scientifically useful concept of corroboration could look like.

1 Introduction. Motivating the concept of corroboration

According to influential accounts of scientific method, scientific knowledge grows by repeatedly testing our best hypotheses (e.g., Popper 1934/2002; Mayo 1996). Nowadays, hypothesis tests have acquired a predominant role in scientific reasoning and are a crucial part of publication requirements. The most frequent form of scientific inference is the *null hypothesis significance test (NHST)*: it tests a precise hypothesis h_0—the "null" or default hypothesis—against an unspecific alternative h_1. In the most common form of NHST, the null hypothesis posits a precise value for a real-valued parameter θ ($h_0 : \theta = \theta_0$), while the alternative ($h_1 : \theta \neq \theta_0$) is a disjunction of uncountably many precise hypotheses (e.g., Neyman and Pearson 1933; Fisher 1956). The null denotes an absent or negligible effect (e.g., a new medical drug is not better than a placebo treatment) whereas the alternative stands for a sizeable effect. NHST are applied across all domains of science, but are especially prominent in psychology and medicine.

Despite their popularity in scientific inference, the philosophical foundations of NHST are shaky at best. NHST are used for quantifying evidence that the data accumulate *against* the null hypothesis. When this level of evidence is high enough, i.e., greater than a prespecified significance threshold, the null hypothesis is *rejected*. However, there is barely any methodological guidance on how to interpret a non-significant result, that is, a result where we *fail to reject the null hypothesis*. Statistics textbooks (e.g., Chase and Brown 2000; Wasserman 2004) restrict themselves to a purely negative interpretation: failure to reject the null means failure to demonstrate a statistically significant phenomenon. This does not address a crucial question in scientific reasoning: Do the results *corroborate* the null hypothesis? Should we *prefer* the null hypothesis to the alternative hypotheses and preliminarily *accept* it? Whenever the null hypothesis is of substantial scientific interest (e.g., independence of two variables in a causal model), such judgments are urgently required.

Explicating degree of corroboration is thus central for a sound interpretation of NHST. How should we explicate it? Karl R. Popper is one of the very few philosophers engaging in this business. He proposes the following characterization:

> By the degree of corroboration of a theory I mean a concise report evaluating the state (at a certain time t) of the critical discussion of a theory, with respect to the way it solves its problems; its *degree of testability*; the *severity of tests* it has undergone; and *the way it has stood up to these tests*.

> Corroboration (or degree of corroboration) is thus an evaluating report of past performance. Like preference, it is essentially comparative. (Popper 1979, 18, my emphasis. See also Popper 1934/2002, 248.)

In Popper's view, corroboration judgments positively appraise the performance of the null hypothesis in a severe test, rather than just stating the failure to find significant evidence *against* it. Notably, high degrees of corroboration need not guide us to the truth (Popper 1979, 21). Instead, the function of corroboration is *comparative* and *pragmatic*: it guides our practical preferences over competing hypotheses, e.g., the choice of the hypothesis on which we base the next experiment (cf. Popper 1934/2002, 416). This is exactly what most statistically working scientists are after when testing a complex set of hypotheses.

Explicating degree of corroboration might thus help to elucidate the value of hypothesis tests in science. Because of the well-known shortcomings of NHST and their practical misuse, it has been suggested that the entire business of hypothesis testing should be abandoned and be replaced by an estimation-centered perspective (Schmidt and al. 1997; Cumming 2015). Sound corroboration judgments may help to respond to this challenge and lead to more nuanced interpretations of hypothesis tests. Especially in classical testing problems like model selection, inference about causal nets, or decisions whether or not to publish a scientific finding, a reliable measure of degree of corroboration may improve scientific reasoning. More generally, a measure of degree of corroboration might revive a critical rationalist epistemology of science, by showing how hypothesis tests increase scientific knowledge (e.g., Rowbottom 2011). In that context, it is notable that neither philosophers nor statisticians have found an adequate explication of degree of corroboration, and that past efforts have been met with devastating criticism (Díez 2011; Rowbottom 2013).

This situation prompts the question of what has been going wrong with the concept of corroboration. My paper answers this question by claiming that the standard formal framework for explicating degree of corroboration does not square well with the task of that concept in scientific reasoning. I will defend this claim by means of two mathematical impossibility results. Broadly speaking, I demonstrate the impossibility of any probabilistic measure of corroboration that is based on both the testability of the hypothesis, and its statistical relevance for the observed evidence. These are, however, the principal virtues that Popper and his successors wanted to capture in a measure of corroboration.

Based on the results of our analysis, we conclude that it is necessary to develop a different framework for explicating degree of corroboration. This will be less Popperian than the original proposals, but closer to actual scientific reasoning. In particular, we hypothesize that an adequate explication of degree of corroboration should be sensitive to the way the alternative hypotheses are partitioned in a scientific inference problem. Spelling out this proposal in detail will be left to future work, though.

The paper is structured as follows. Section 2 briefly presents Popper's characterization of an adequate measure of degree of corroboration. Section 3 is the core if the paper: it develops plausible adequacy criteria for degree of corroboration in a statistical relevance framework and demonstrates that no measure of corroboration can satisfy them all. The final Section 4 discusses our findings and explores ways out of the dilemma created by the impossibility results.

2 Popper's Measure of Degree of Corroboration

Popper's first writings on degree of corroboration, in Chapter 10 of "The Logic of Scientific Discovery" (1934/2002), do not engage in a quantitative explication. Apparently, this task is deferred to a scientist's common sense. However, this move makes the entire concept of corroboration vulnerable to the charge of subjectivism: without a quantitative criterion, it is not clear which corroboration judgments are sound and which aren't (Good 1968, 136). Especially if we aim at gaining *objective knowledge* from hypothesis tests, we need a precise explication of degree of corroboration.

Popper faces this challenge in a couple of *BJPS* articles (Popper 1954, 1957, 1958) that form, together with a short introduction, appendix ix of "The Logic of Scientific Discovery". In these articles, Popper develops and defends a measure of degree of corroboration. Popper argues that this measure cannot be a probability in the sense of Carnap (1950), i.e., the plausibility of the tested theory (or hypothesis) conditional on the observed evidence:

> [...] the probability of a statement [...] simply does not express an appraisal of the severity of the tests a theory has passed, of the manner in which it has passed these tests. (Popper 1934/2002, 411)

In particular, logical content and informativity contribute to the testability of a theory, and therefore also to its degree of corroboration:

> The main reason for this is that the *content* of a theory—which is the same as its *improbability*—determines its *testability* and *corroborability*. (ibid., original emphasis)

Instead of co-varying with probability, corroboration should be sensitive to the logical content of a theory. Here we see a major difference to Bayesian measures of evidential support. A reason could be that corroboration is also supposed our judgments of *acceptance*, where it is often required that good theories be informative (see the discussions in Hempel 1960; Levi 1963; Huber 2005, 2008). Indeed, Popper confirms that scientific theory assessment pursues both goals at once:

> Science does not aim, primarily, at high probabilities. It aims at a *high informative content*, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or little. (Popper 1934/2002, 416, original emphasis)

Such a characterization of corroboration is attractive because it would amalgamate two crucial cognitive values in theory assessment: high informative content and empirical support. Also in NHST, both values play a role since a precise hypothesis (the null) is tested against a continuum of alternatives. However, this paper shows that such a tradeoff is unattainable if further reasonable assumptions are made.

Let us now have a look at how Popper characterizes degree of corroboration. Transcribed to modern notation, Popper assumes that evidence e and hypothesis h are among the closed sentences $\mathfrak{L}$ of a first-order language L. A corroboration measure is then described by a function $c : \mathfrak{L}^2 \times \mathfrak{P} \to \mathbb{R}$, where $\mathfrak{P}$ is the set of probability measures on the σ-algebra generated by $\mathfrak{L}$. This function assigns a real-valued degree of corroboration $c(h, e)$ to any pair of sentences in $\mathfrak{L}$, together with a probability measure $p(\cdot)$. This measure may be interpreted as a function of the logical structure of L, but also as objective chance or degree of belief—our discussion is independent of this point. For the sake of simplicity, we will omit reference to background assumptions and assume that they are implicit in the probability function $p(\cdot)$.

Then a set of adequacy criteria is specified.

I $\quad c(h, e) >/=/< 0 \qquad$ if and only if $\qquad p(e|h) >/=/< p(e|\neg h)$.

This is a classical *statistical relevance condition*: e corroborates h just in case e is more expected under h than under $\neg h$. This condition is also in line with Popper's remark that corroboration is, like preference, essentially contrastive (Popper 1979, 18).

II $-1 = c(h, \neg h) \leq c(h, e) \leq c(h, h) \leq 1$.

III $c(h, h) = 1 - p(h)$.

IV If $e \models h$ then $c(h, e) = 1 - p(h)$.

V If $e \models \neg h$ then $c(h, e) = -1$.

These conditions determine under which conditions the measure of corroboration takes its extremal values. Minimal degree of corroboration is obtained if the evidence refutes the hypothesis (V). Conversely, the most corroborating piece of evidence e is a verification of h. In that case, degree of corroboration is equal to the *improbability* of h (II, III, IV), which is supposed to express the informativity, testability and logical content of h.[1] See Popper (1934/2002, 268–269), Popper (1963, 385–387), Rowbottom (2013, 741–744), and the above discussion. Assigning a corroboration bonus to highly informative and testable hypotheses also fits well into a critical rationalist picture about goals and method of science.

VI $c(h, e) \geq 0$ increases with the power of h to explain e.

VII If $p(h) = p(h')$, then $c(h, e) > c(h', e')$ if and only if $p(h|e) > p(h'|e')$.

These conditions reiterate the statistical relevance rationale from condition I, and make it more precise. Regarding condition VI, Popper (1934/2002, 416) defines explanatory power according to the formula $\mathcal{E}(e, h) = (p(e|h) - p(e))/(p(e|h) + p(e))$, another measure of the statistical relevance between e and h. But the details need not bother us here. Condition VII states that corroboration essentially co-varies with posterior probability whenever two hypotheses are equiprobable at first. In that case, posterior probability is a good indicator of statistical relevance.

VIII If $h \models e$, then

a) $c(h, e) \geq 0$;

b) $c(h, e)$ is an increasing function of $1 - p(e)$;

c) $c(h, e)$ is an increasing function of $p(h)$.

IX If $\neg h$ is consistent and $\neg h \models e$, then

[1] This is especially plausible in Carnap's logical interpretation of probability, which Popper adopts for $p(h)$. But it also makes sense for a subjective Bayesian interpretation.

a) $c(h,e) \leq 0$;

b) $c(h,e)$ is an increasing function of $p(e)$;

c) $c(h,e)$ is an increasing function of $p(h)$.

Condition VIII demands that corroboration gained from a successful deductive prediction co-vary with the informativity of the evidence and the prior probability of the hypothesis. Condition IX mirrors this requirement for the case $\neg h \models e$. These conditions can be motivated from the idea that if $h \models e$, then corroboration should not automatically transfer to hypotheses $h \wedge h'$ that contain an "irrelevant conjunct" h' which has not yet been tested. See the next section for more detailed discussion of this point.

Popper then proposes the corroboration measure $c_P(h,e)$ which satisfies all of his constraints:

$$c_P(h,e) = \frac{p(e|h) - p(e)}{p(e|h) + p(e) - p(e|h)\, p(h)}. \tag{1}$$

But we can easily see that an essential motivation behind a measure of degree of corroboration is not satisfied. $c_P(h,e)$ is an increasing function of $p(h)$ for all values of $p(e|h)$ and $p(e)$. Hence, the informativity of the tested hypothesis never contributes to its degree of corroboration. This violates Popper's informal characterization of the concept and does not square well with the practice of NHST. The only exception is the case $p(h|e) = 1$, as expressed in IV, but then we are arguably not in need of a measure of corroboration: h has been proved conclusively.[2] We shall now see that this problem does not dissolve when moving from Popper's proposal to a broader class of corroboration meaures.

3 The Impossibility Results

Popper's nine adequacy conditions are quite specific requirements and too strong for the purpose of a general analysis. We will therefore weaken them and retain only such adequacy conditions that we consider indispensable for an analysis of corroboration. We then show two impossibility results for corroboration measures that (i) are built on the notion of statistical relevance between e and h; and (ii) preserve the intuition that corroboration should not co-vary with prior probability, but also be sensitive to the informativity (and testability) of the tested hypothesis.

[2]See Díez (2011) for a more detailed analysis of why Popper's own explication is at odds with the general tenets of critical rationalism.

I would like to begin with a condition which is mainly representational in nature and has proved its mettle in formal epistemology (cf. Schupbach and Sprenger 2011; Crupi 2014):

Formality There exists a function $f : [0,1]^3 \to \mathbb{R}$ such that for all $e, h \in \mathfrak{L}$ and $p(\cdot) \in \mathfrak{P}$,

$$c(h,e) = f(p(e|h), p(e), p(h)).$$

This condition relates degree of corroboration depends to the joint probability distribution of e and h. The three arguments of f determine that distribution in all non-degenerate cases, and they are the same quantities that figure in Popper's measure of corroboration c_P. This makes comparisons easier. In practice, `Formality` means that two scientists who agree about all relevant probabilities will make the same corroboration judgments.

Now let's move to the substantial conditions. A natural condition for degree of corroboration is the following, familiar constraint:

Weak Law of Likelihood (WLL) For mutually exclusive hypotheses $h_1, h_2 \in \mathfrak{L}$, $e \in \mathfrak{L}$ and $p(\cdot) \in \mathfrak{P}$, if

$$p(e|h_1) \geq p(e|h_2) \qquad \text{and} \qquad p(e|\neg h_1) \leq p(e|\neg h_2) \tag{2}$$

with one inequality being strict, then $c(h_1, e) > c(h_2, e)$.

The WLL has been defended as capturing a "core message of Bayes' Theorem" (Joyce 2008): if h_1 predicts e better than h_2, and $\neg h_2$ predicts e better than $\neg h_1$, then e favors h_1 over h_2. Since WLL is phrased in terms of predictive performance, it is even more compelling for corroboration than for evidential support. After all, $p(e| \pm h1)$ and $p(e| \pm h2)$ measure how well h_1 and h_2 have stood up to a test with outcome e. The version given here is in one sense weaker and in one sense stronger than Joyce's original formulation: it is stronger because only one inequality has to be strict (see also Brössel 2013: 395–396); it is weaker because the WLL has been restricted to mutually exclusive hypotheses, where our intuitions tend to be more reliable.

Another condition deals with irrelevant evidence:

Screened-Off Evidence Let $e_1, e_2, h \in \mathfrak{L}$ and $p \in \mathfrak{P}$. If e_2 is probabilistically independent of e_1, h, and $e_1 \wedge h$ and $p(e_2) > 0$. Then $c(h, e_1) = c(h, e_1 \wedge e_2)$.

This condition prominently figures in several explications of evidential support and explanatory power (e.g., Kemeny and Oppenheim 1952; Schupbach and Sprenger 2011). But it is also very sensible with respect to degree of corroboration. In an experiment where h has been tested and (relevant) evidence e_1 has been observed, completely irrelevant extra evidence ($e_2 \perp\!\!\!\perp e_1, h, e_1 \wedge h$) should not change the evaluation of the results. Imagine, for example, that a scientist tests the hypothesis that a high pitch facilitates voice recognition. As her university is interested in improving the planning of lab experiments, the scientist also collects data on when participants drop in, which days of the week are busy, which ones are quiet, etc. Plausibly, these data satisfy the independence conditions of `Screened-Off Evidence`. But equally plausibly, they do not influence the degree of corroboration of the hypothesis under investigation.

The next adequacy condition is motivated by the problem of irrelevant conjunctions (e.g., Fitelson 2002; Hawthorne and Fitelson 2004). Assume that a hypothesis h, such as General Theory of Relativity (GTR), logically implies a phenomenon e, such as the perihelion shift of Mercury. This observation corroborates GTR: logical implication is a special case of statistical relevance.

However, once we add an utterly irrelevant proposition h' = "the chicken came before the egg" to the hypothesis, it seems that e corroborates $h \wedge h'$—the *conjunction* of GTR and the chicken-egg hypothesis—not more than h, if at all. After all, h' was in no way tested by the observations we made. It has no record of past performance to which we could appeal. This motivates the following constraint:

Irrelevant Conjunctions Assume the following conditions on $h, h', e \in \mathfrak{L}$ and $p \in \mathfrak{P}$ are satisfied:

[1] h and h' are consistent and $p(h \wedge h') < p(h)$;

[2] $p(e) \in (0, 1)$;

[3] $h \models e$;

[4] $p(e|h') = p(e)$.

Then it is always the case that $c(h \wedge h', e) \leq c(h, e)$.

This requirement states that for any non-trivial hypothesis h' that is consistent with h ([1]) and irrelevant for e ([4]), $h \wedge h'$ is corroborated no more than h whenever h non-trivially entails e ([2], [3]). Indeed, it would be strange if corroboration could be increased "for free" by attaching irrelevant conjunctions. Plausibly, this requirement

may be strengthened to a strict inequality, but for our purposes, the weaker formulation is sufficient.

Interestingly, the preceding adequacy conditions can be derived from Popper's original adequacy conditions (all proofs are given in the appendix):

Theorem 1 The following statements are true:

- Popper's condition VII implies `Weak Law of Likelihood` for the case of equiprobable hypotheses.
- Popper's condition VII implies `Screened-Off Evidence`.
- Popper's condition VIIIc implies `Irrelevant Conjunctions`.

This shows that our adequacy conditions are motivated in the right way: they are weaker versions of the criteria that Popper set up in "The Logic of Scientific Discovery". We can thus be confident that our formal analysis of corroboration is on target and that our adequacy conditions do not track a different, incompatible concept.

However, unlike evidential support, corroboration contains an element of severe testing: the hypothesis should run a risk of being falsified, and high informativity and logical content contribute to this goal. Highly corroborated hypotheses are *informative* propositions, well-backed by the evidence (cf. Popper's quote on page 6 and conditions III and IV). This motivates the following desideratum:

Weak Informativity Degree of corroboration $c(h, e)$ does not generally increase with the probability of h. That is, there are $h, h', e \in \mathfrak{L}$ and $p \in \mathfrak{P}$ such that

(1) $p(e|h) = p(e|h') > p(e)$;
(2) $1/2 \geq p(h) > p(h')$;
(3) $c(h, e) \leq c(h', e)$.

The intuition behind `Weak Informativity` can also be expressed as follows: corroboration does not, in the first place, assess the probability of a hypothesis; therefore $c(h, e)$ should not *always* increase with the probability of h. To this, the following condition—`Strong Informativity`—adds that low probability/high logical content can in principle be corroboration-conducive. Note that the requirement $1/2 \geq p(h), p(h')$ is purely technical and philosophically innocuous.

Strong Informativity The informativity/logical content of a proposition can increase degree of corroboration, ceteris paribus. That is, there are $h, h', e \in \mathfrak{L}$ and $p \in \mathfrak{P}$ such that

(1) $p(e|h) = p(e|h') > p(e)$;

(2) $1/2 \geq p(h) > p(h')$;

(3) $c(h, e) \leq c(h', e)$.

To my mind, any account of corroboration that denies these properties has stripped itself of its distinctive features with respect to evidential support, or statistical relevance more generally. At the very least, the *Popperian* characaterization of corroboration as capturing both predictive success and testability would have to be abandoned, and links with NHST would have to be loosened.

Now, we demonstrate that the listed adequacy conditions are incompatible with each other. First, it is a consequence of `Weak Law of Likelihood` that corroboration increases with the probability of a hypothesis. This clashes directly with `Strong/Weak Informativity`:

Theorem 2 No measure of corroboration $c(h, e)$ constructed according to `Formality` can satisfy `Weak Law of Likelihood` and `Weak/Strong Informativity` at the same time.

Since `Formality` is a purely representational condition, this result means that `Weak Law of Likelihood` and `Weak/Strong Informativity` pull into different directions: the first condition emphasizes the predictive performance of the tested hypothesis, the second its logical strength. It is perhaps surprising that these two conditions are already incompatible, since it is a popular tenet of critical rationalism that informative hypotheses are also more valuable predictively.

Second, and even more surprisingly, `Strong Informativity` clashes with `Irrelevant Conjunctions` and `Screened-Off Evidence`:

Theorem 3 No measure of corroboration $c(h, e)$ constructed according to `Formality` can satisfy `Screened-Off Evidence`, `Irrelevant Conjunctions` and `Strong Informativity` at the same time.

Thus, the intuition behind `Strong/Weak Informativity` cannot be satisfied if other plausible adequacy constraints on degree of corroboration are accepted. In particular, if a measure of corroboration is insensitive to irrelevant evidence and does not reward adding irrelevant conjunctions, then it cannot give any bonus to informative hypotheses. The less informative and testable a hypothesis is, the higher its degree of corroboration, ceteris paribus.

Finally, the result of Theorem 3 can be extended to `Weak Informativity` if we make the assumption that irrelevant conjunctions *dilute* the degree of corroboration, rather than not increasing it (proof omitted). See also the corresponding remark on p. 10, in the motivation of `Irrelevant Conjunctions`.

Note that these results are meaningful even if somebody is not interested in the project of explicating Popperian corroboration (e.g., because she is a radical subjective Bayesian). Some of the above adequacy conditions have been proposed for measures of evidential support or explanatory power; others could be potentially interesting in these contexts. For instance, Brössel (2013) has recently discussed the condition `Logicality`, which resembles our `Strong/Weak Informativity`. Hence, our results also make sense in the framework of Bayesian Confirmation Theory, as indicating the impossibility of probabilistic measures that caputre informativity and statistical relevance at the same time.

All this does not yet show that explicating degree of corroboration is a futile project. Rather, it reveals a fundamental and insoluble tension between the two main contributing factors of corroboration that Popper identifies (see the quote on p. 6): statistical relevance and testability. `Weak Law of Likelihood`, `Screened-Off Evidence` and `Irrelevant Conjunctions` all speak to the statistical relevance intuition, whereas `Strong/Weak Informativity` rewards high logical content and testability. That it is impossible to satisfy minimal subsets of these plausible conditions sheds doubts on the prospects for explicating corroboration in a statistical relevance framework. However, before we prematurely draw pessimistic conclusions, let us revisit the available options.

4 Discussion

In this paper, we have first demonstrated the urgency of searching for an adequate probabilistic measure of corroboration. This has been motivated by the lack of guidance on the interpretation of non-significant results in statistical hypothesis tests (NHST). We have then explored Popper's idea that a measure of corroboration should capture both the statistical relevance of evidence and hypothesis, and the testability of the hypothesis. To this end, we have set up a set of plausible conditions that are weaker than Popper's original claims (Theorem 1).

However, it turns out that these criteria cannot be jointly satisfied. The pre-theoretic concept of corroboration is overloaded with desiderata that point into different direc-

tions and create insoluble tensions (Theorem 2 and 3). This leaves us with four options: (i) to reject one of the (substantial) adequacy conditions; (ii) to split up degree of corroboration into different sub-concepts, as it happened for evidential support; (iii) to conclude that the explication of degree of corroboration is hopeless and not worthy of further pursuit, and (iv) to blame the *representational framework* that has been used for explicating degree of corroboration, and to look for explications in a different style.

Option (i) would come down to either giving up `Weak Law of Likelihood`, `Screened-Off Evidence`, `Irrelevant Conjunctions` or `Strong/Weak Informativity`. But each of these adequacy conditions for degree of corroboration has been carefully motivated in the preceding section. Such a step would therefore appear arbitrary and unsatisfactory.

Option (ii) amounts to endorsing pluralism for degree of corroboration. The model case for this option are probabilistic analyses of evidential support: some measures, like $d(h,e) = p(h|e) - p(h)$ capture the *boost in degree of belief* in h provided by e, while others, like $l(h,e) = p(e|h)/p(e|\neg h)$, aim at the *discriminatory power* of e with respect to h and $\neg h$. However, it is not clear what similarly interesting subconcepts could look like for degree of corroboration. Right now, this option does not appear to be viable.

Neither does the pessimistic option (iii) have much appeal, unless convincing reasons are given why scientists can dispense with the concept of corroboration, and hypothesis testing in general.

This leaves us with option (iv). Here two strategies are possible. One of them endorses Bayes factors or another statistical relevance measure as measures of corroboration, giving up the informativity intuition. This has the advantage of relating corroboration to a bunch of statistical and philosophical literature (e.g., Fitelson 1999), but it comes at the price of stripping the concept of its defining characteristics. It might then become redundant with respect to evidential support.

Also, statistical relevance measures generally depend on $p(e|\neg h)$, either explicitly (like the l-measure or Bayes factors) or via the calculation of $p(e)$ and $p(h|e)$. This creates a variety of problems. Consider, for example, the case of a Binomial model where we test the null hypothesis $h_0 : \theta = 0.5$ against the alternative $h_1 : \theta \neq 0.5$. If the observed relative frequency of successes is close to 0.5, e.g., $\bar{x} = 0.53$, the degree corroboration of the null hypothesis does not seem to depend on the likelihoods $p(\bar{x}|\theta)$ for very large and very small values of θ. But for statistical relevance measures, this conclusion is inevitable since $p(x|\theta \neq \theta_0) = \int_0^1 p(\theta)p(x|\theta)d\theta$.

Therefore we might consider the second strategy: to abandon the entire statistical

relevance framework. Perhaps it is neither necessary nor sufficient to base a corroboration judgment on the joint probability distribution of h and e? As noted above, statistical relevance measures of corroboration compare the merits of h with the merits of $\neg h$, defined as the *aggregate* of alternatives to h. However, a comparison to such an aggregate does not make much sense in many NHST contexts where we deal with a multitude of distinct alternatives h_i, $i \in \mathbb{N}$. Perhaps corroboration judgments should be made with respect to the best-performing alternative in the hypothesis space, and not with respect to *all* possible alternatives.

This suggests that we might develop explications of degree of corroboration in a framework with *many distinct alternatives* to the tested hypothesis h. As a consequence, `Formality` would have to be dropped and degree of corroboration would become partition-relative: testing h with alternative $\neg h$ can lead to different corroboration judgments than testing h with alternatives $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ even if $\neg h = \bigwedge_{1 \leq i \leq n} h_i$.[3]

The work of spelling out this approach in detail and providing axiomatic foundations is left to future work. Note that this paper's main result is negative and constructive at the same time: it shows why there can be no measure of corroboration that fits Popper's informal description, or more generally, that amalgamates statistical relevance with informativity and testability. At the same time, the paper motivates why we have to expand our mathematical framework for explicating degree of corroboration, and which type of explications could prove useful for science and philosophy at the same time.

[3]Such an approach has been anticipated by I.J. Good (1960, 1968). However, Good opts for a vector-valued measure of degree of corroboration, which is, for many reasons, unhelpful in scientific practice.

A Proofs

Proof of Theorem 1: Assume $p(h_1) = p(h_2)$. We distinguish two jointly exhaustive cases in which `WLL` may apply:

Case 1: $p(e|h_1) > p(e|h_2)$ Case 2: $p(e|h_1) = p(e|h_2)$ and $p(e|\neg h_1) < p(e|\neg h_2)$.

For the first case, the proof is simple in virtue of the inequality

$$p(h_1|e) = p(h_1)\frac{p(e|h_1)}{p(e)} > p(h_2)\frac{p(e|h_2)}{p(e)} = p(h_2|e).$$

Then, VII guarantees that $c(h_1, e) > c(h_2, e)$.

For the second case, let $x := p(e|h_1) = p(e|h_2)$ and $y := p(h_1) = p(h_2)$. We know that

$$\begin{aligned} p(e|\neg h_1) &= \frac{1}{1-p(h_1)}\left[p(e|h_2)p(h_2) + p(e|\neg h_1\neg h_2)p(\neg h_1\neg h_2)\right] \\ &= \frac{1}{1-y}(xy + p(e|\neg h_1\neg h_2)p(\neg h_1\neg h_2)) \\ p(e|\neg h_2) &= \frac{1}{1-p(h_2)}\left[p(e|h_1)p(h_1) + p(e|\neg h_1\neg h_2)p(\neg h_1\neg h_2)\right] \\ &= \frac{1}{1-y}(xy + p(e|\neg h_1\neg h_2)p(\neg h_1\neg h_2)). \end{aligned}$$

Hence, $p(e|\neg h_1) = p(e|\neg h_2)$. On the other hand, we have assumed that $p(e|\neg h_1) < p(e|\neg h_2)$. This shows that the second case can never occur and may be dismissed.

We now prove the second implication, that is, VII $\Rightarrow$ `Screened-Off Evidence`. To this end, we rewrite VII as

VII If $p(h) = p(h')$, then $c(h, e) \leq c(h', e')$ if and only if $p(h|e) \leq p(h'|e')$.

By assuming $h = h'$, it is easy to see that VII implies

VII' If $p(h|e) = p(h|e')$, then $c(h, e) = c(h, e')$.

The reason is simple: If $p(h|e) = p(h|e')$, then also $p(h|e) \leq p(h|e')$ and the '$\Leftarrow$' direction of VII implies $c(h, e) \leq c(h, e')$, where h has been substituted for h'. Now we

repeat the same trick with the premise $p(h|e') \leq p(h|e)$ and we obtain $c(h,e') \leq c(h,e)$. Taking both inequalities together yields the conclusion $c(h,e) = c(h,e')$ and thereby VII′.

Notice that under the conditions of `Screened-Off Evidence`, $p(h|e_1 \wedge e_2) = p(h|e_1)$. This is so because

$$p(h|e_1 \wedge e_2) = p(h)\frac{p(e_1 \wedge e_2|h)}{p(e_1 \wedge e_2)} = p(h)\frac{p(e_1|h)\,p(e_2)}{p(e_1)\,p(e_2)} = p(h)\frac{p(e_1|h)}{p(e_1)} = p(h|e_1).$$

Hence, we can apply VII′ to the case of `Screened-Off Evidence`, with $e := e_1$ and $e' := e_1 \wedge e_2$. This implies

$$c(h, e_1 \wedge e_2) = c(h, e_1),$$

completing the proof.

Finally, we have the implication VIIIc $\Rightarrow$ `Irrelevant Conjunctions`. Let for $h, h', e \in \mathfrak{L}$ and $p \in \mathfrak{P}$ the conditions of `Irrelevant Conjunctions` ([1] to [4]) be satisfied. Since $h \models e$, VIIIc implies that $c(h,e)$ and $c(h \wedge h', e)$ are increasing functions of the probability of the tested hypothesis ($p(h)$ and $p(h \wedge h')$, respectively). But by assumption, we have $p(h \wedge h') < p(h)$. Hence, it follows that $c(h \wedge h', e) < c(h,e)$. $\square$

Proof of Theorem 2: By `Weak Informativity`, there are $x > y$ and $z > z'$ with $z + z' < 1$:

$$f(x,y,z) \leq f(x,y,z').$$

Choose a probability function $p(\cdot)$ such that $p(h_1) = z$, $p(h_2) = z'$, $p(h_1 \wedge h_2) = 0$, $p(e|h_1) = p(e|h_2) = x$, $p(e) = y$. This is always possible because it was assumed that $z + z' < 1$. Then it is straightforward to show that

$$\begin{aligned}
p(e|\neg h_1) &= \frac{1}{1-p(h_1)}\left[p(e|h_2)p(h_2) + p(e|\neg h_1 \neg h_2)p(\neg h_1 \neg h_2)\right] \\
&= \frac{1}{1-p(h_1)}\left[p(e|h_1)p(h_2) + p(e|\neg h_1 \neg h_2)p(\neg h_1 \neg h_2)\right] \\
p(e|\neg h_2) &= \frac{1}{1-p(h_2)}\left[p(e|h_1)p(h_1) + p(e|\neg h_1 \neg h_2)p(\neg h_1 \neg h_2)\right]
\end{aligned}$$

because by assumption, $p(e|h_1) = p(e|h_2)$. From this we can infer

$$
\begin{aligned}
& p(e|\neg h_1) - p(e|\neg h_2) \\
= \;& \frac{p(e|h_1)\, p(h_2)}{1 - p(h_1)} + \frac{p(e|\neg h_1 \neg h_2)\, p(\neg h_1 \neg h_2)}{1 - p(h_1)} - \frac{p(e|h_1)\, p(h_1)}{1 - p(h_2)} - \frac{p(e|\neg h_1 \neg h_2)\, p(\neg h_1 \neg h_2)}{1 - p(h_2)} \\
= \;& p(e|h_1) \left[\frac{p(h_2)}{1 - p(h_1)} - \frac{p(h_1)}{1 - p(h_2)} \right] + p(e|\neg h_1 \neg h_2)(1 - p(h_1) - p(h_2)) \\
& \cdot \left[\frac{1}{1 - p(h_1)} - \frac{1}{1 - p(h_2)} \right] \\
= \;& p(e|h_1) \frac{p(h_2) - p(h_2)^2 - p(h_1) + p(h_1)^2}{(1 - p(h_1))\,(1 - p(h_2))} + p(e|\neg h_1 \neg h_2) \\
& \cdot (1 - p(h_1) - p(h_2)) \, \frac{p(h_1) - p(h_2)}{(1 - p(h_1))\,(1 - p(h_2))} \\
= \;& p(e|h_1) \frac{(p(h_1) - p(h_2)) \cdot (p(h_1) + p(h_2) - 1)}{(1 - p(h_1))\,(1 - p(h_2))} + p(e|\neg h_1 \neg h_2) \\
& \cdot (1 - p(h_1) - p(h_2)) \, \frac{p(h_1) - p(h_2)}{(1 - p(h_1))\,(1 - p(h_2))} \\
= \;& \frac{(p(h_1) - p(h_2)) \cdot (p(h_1) + p(h_2) - 1)}{(1 - p(h_1))\,(1 - p(h_2))} \, (p(e|h_1) - p(e|\neg h_1 \neg h_2)).
\end{aligned}
$$

If we look at the signs of the involved factors, we notice first that $p(h_1) = z > z' = p(h_2)$ and $p(h_1) + p(h_2) - 1 = z + z' - 1 < 0$. Then we observe that h_1 and h_2 were disjoint and that $p(e|h_1)$ and $p(e|h_2)$ are both greater than $p(e)$, implying $p(e|h_1) > p(e|\neg h_1 \neg h_2)$. Taken together, we can then conclude

$$p(e|\neg h_1) - p(e|\neg h_2) < 0.$$

Hence, the conditions for applying `Weak Law of Likelihood` are satisfied:

$$f(x, y, z) = c(h_1, e) > c(h_2, e) = f(x, y, z'),$$

in contradiction with the inequality $f(x, y, z) \leq f(x, y, z')$ that we got from `Weak Informativity`. □

Lemma 1 Any measure of corroboration $c : \mathfrak{L}^2 \times \mathfrak{P} \to \mathbb{R}$ that satisfies `Screened-Off Evidence` and `Formality` also satisfies the equality

$$f(ax, ay, z) = f(x, y, z) \tag{3}$$

for $x > y > 0$, $z > 0$ and $0 < a \leq 1$.

Proof of Lemma 1: For any $x > y > 0$, $z > 0$ and $0 < a \leq 1$, we can choose sentences $h, e_1, e_2 \in \mathfrak{L}$ and a probability function $p(\cdot) \in \mathfrak{P}$ such that

$$\begin{aligned} a &:= p(e_2) & p(e_2 h) &= p(e_2)p(h) \\ x &:= p(e_1|h) & p(e_1 \wedge e_2) &= p(e_2)p(e_1) \\ y &:= p(e_1) & p(e_1 \wedge e_2|h) &= p(e_2)p(e_1|h) \\ z &:= p(h). \end{aligned}$$

Since our choice of p is not restricted, this is always possible. Now, the conditions of `Screened-Off Evidence` are satisfied, and it follows that $c(h, e_1 \wedge e_2) = c(h, e_1)$. By `Formality`, we can also derive the equalities

$$\begin{aligned} c(h, e_1 \wedge e_2) &= f(p(e_1 \wedge e_2|h), p(e_1 \wedge e_2), p(h)) = f(p(e_2)p(e_1|h), p(e_2)p(e_1), p(h)) \\ &= f(ax, ay, z) \\ c(h, e_1) &= f(x, y, z). \end{aligned}$$

Taking all these equalities together delivers the desired result:

$$f(ax, ay, z) = c(h, e_1 \wedge e_2) = c(h, e_1) = f(x, y, z).$$

□

Proof of Theorem 3: Choose sentences h_1, h_2, $e \in \mathfrak{L}$ and a probability function $p(\cdot) \in \mathfrak{P}$ such that the conditions of `Strong Informativity` are satisfied:

(1) $p(e|h_1) = p(e|h_2) > p(e)$;

(2) $1/2 \geq p(h_1) > p(h_2)$;

(3) $c(h_1, e) < c(h_2, e)$.

Writing $x := p(e|h_1) = p(e|h_2)$, $y := p(e)$, $z = p(h)$ and $z' := p(h')$, we then obtain

$$f(x, y, z) = c(h_1, e) < c(h_2, e) = f(x, y, z'). \tag{4}$$

Since $c(h, e)$ satisfies `Formality` and `Screened-Off Evidence`, by Lemma 1 it also satisfies the equality

$$f(ax, ay, z) = f(x, y, z)$$

for $x > y > 0$, $z > 0$ and $0 < a \leq 1$. With $a := x$, we now obtain

$$f(1, y/x, z) = f(x, y, z) \qquad\qquad f(1, y/x, z') = f(x, y, z').$$

It then follows from inequality (4) and the above equalities that

$$f(1, y/x, z) < f(1, y/x, z') \tag{5}$$

for these specific values of x, y, z and z'.

We can now find sentences h, h', e' and a probability function $p'(\cdot)$ such that the conditions of `Irrelevant Conjunctions` are satisfied, and moreover $p'(h) = z$, $p'(h \wedge h') = z'$, $p'(e') = y/x$. This implies $c(h \wedge h', e') \leq c(h, e')$. By `Formality`, this also implies

$$f(1, y/x, z) \geq f(1, y/x, z').$$

However, this inequality contradicts equation (5) that we have shown before. Hence, the theorem is proven. □

References

Brössel, P. (2013): "The Problem of Measure Sensitivity Redux", *Philosophy of Science* 80, 378–397.

Carnap, R. (1950): *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

Chase, W., and F. Brown (2000): *General Statistics*. New York: Wiley.

Crupi, V. (2014): "Confirmation", in: E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/entries/confirmation/`, retrieved on January 30, 2015.

Cumming, G. (2015): "The New Statistics", forthcoming in *Psychological Science*.

Díez, J. (2011): "On Popper's strong inductivism (or strongly inconsistent anti-inductivism)", *Studies in the History and Philosophy of Science A* 42, 105–116.

Fisher, R.A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Fitelson, Branden (1999): "The plurality of Bayesian measures of confirmation and the problem of measure sensitivity"", *Philosophy of Science (Proceedings)* 66, S362–S378.

Fitelson, B. (2002): "Putting the Irrelevance Back Into the Problem of Irrelevant Conjunction", *Philosophy of Science* 69, 611–622.

Good, I.J. (1960): "Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments", *Journal of the Royal Statistical Society B* 22, 319–331.

Good, I.J. (1968): "Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor", *The British Journal for the Philosophy of Science* 19, 123–143.

Hempel, C.G. (1960): "Inductive inconsistencies", *Synthese* 12, 439–469.

Hawthorne, J., and B. Fitelson (2004): "Re-solving Irrelevant Conjunction with Probabilistic Independence", *Philosophy of Science* 71, 505-514.

Howson, C. and P. Urbach (2006): *Scientific Reasoning: The Bayesian Approach*. Third Edition. La Salle: Open Court.

Huber, F. (2005): "What is the Point of Confirmation?", *Philosophy of Science* 72, 1146–1159.

Huber, F. (2008): "Hempel's Logic of Confirmation", *Philosophical Studies* 139, 181–189.

Joyce, J. (2008): "Bayes' Theorem", in: E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/entries/confirmation/`, retrieved on January 30, 2015.

Kemeny, J.G., and P. Oppenheim (1952): "Degrees of factual support", *Philosophy of Science* 19, 307–324.

Levi, I. (1963): "Corroboration and Rules of Acceptance", *The British Journal for the Philosophy of Science* 13, 307–313.

Mayo, D.G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago & London: The University of Chicago Press.

Neyman, J., and E. Pearson (1933): "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society A* 231, 289–337.

Popper, K.R. (1934/2002): *Logik der Forschung*. Berlin: Akademie Verlag. Translated as *The Logic of Scientific Discovery*, 1959. Reprinted in 2002. Routledge: London.

Popper, K.R. (1954): "Degree of Confirmation", *The British Journal for the Philosophy of Science* 5, 143–149.

Popper, K.R. (1957): "A Second Note on Degree of Confirmation", *The British Journal for the Philosophy of Science* 7, 350–353.

Popper, K.R. (1958): "A third note on degree of corroboration or confirmation", *The British Journal for the Philosophy of Science* 8, 294–302.

Popper, K.R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Harper.

Popper, K.R. (1979): *Objective knowledge: an evolutionary approach*. Oxford: Clarendon Press.

Rowbottom, D.P. (2011): *Popper's Critical Rationalism: A Philosophical Investigation*. London: Routledge.

Rowbottom, D.P. (2013): "Popper's Measure of Corroboration and P(h|b)", *The British Journal for the Philosophy of Science* 64, 739–745.

Schmidt, F.L., and J.E. Hunter (1997): "Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data", in: Lisa L. Harlow et al. (eds.), *What if there were no significance tests?*, 37–64. Mahwah/NJ: Erlbaum.

Schupbach, J., and J. Sprenger (2011): "The Logic of Explanatory Power", *Philosophy of Science* 78, 105–127.

Wasserman, L. (2004): *All of Statistics*. New York: Springer.