Confirmation

Gary Hardegree
Department of Philosophy
University of Massachusetts
Amherst, MA 01003

---

1. Hypothesis Testing

In this chapter, we examine a general topic in philosophy of science – hypothesis testing and confirmation. In this context, an *hypothesis*[1] is a proposition proposed for consideration (by the relevant community – e.g., scientists, or philosophers). On the other hand, a *test* of an hypothesis H is an examination of "data" on the basis of which (it is hoped) one can decide whether to accept or reject H.

In thinking about data and hypotheses, the metaphor that we adopt is similar to the one adopted by in our court system. The ideal of justice is often artistically portrayed as a muse called "Lady Justice", the following being a simple graphical example.

The idea then is that, in a court of law, evidence is brought to bear on a case, and it is weighed on the scales of justice. Moreover, the weighing is impartial, which is artistically conveyed by the blindfold.

[1] The prefix 'hypo' means "below" or "beneath", the most common examples being in the word 'hypodermic' – which literally means "below the skin" – and the word 'hypothermia' – which literally means "below heat" or "low heat". The word 'hypothesis' then literally means "below a thesis". What is semantically odd is that one of the definitions of 'thesis' is "a hypothetical proposition", which means that we have a semantic infinite regress. This is not quite right, however, since 'thesis' comes from the Greek *tithenai* [to put]. So an hypothesis is something that is "put below".

The weighing metaphor involves a classification of data into categories and subcategories. First, in reference to a given hypothesis H, we can divide data into two general categories.[2]

(1) data that are *relevant* to H
(2) data that are *irrelevant* to H

The relevant data are in turn sub-divided into two categories.

(2a) data that *favor* H [*positive evidence* for H]
(2b) data that *disfavor* H [*negative evidence* against H]

Returning to the scales of justice, the positive evidence is *weighed against* the negative evidence, and other data are deemed by the judge to be immaterial and irrelevant. As a consequence of the weighing of the evidence, we have three logically possible outcomes.

(1) the positive evidence *outweighs* the negative evidence
(2) the negative evidence *outweighs* the positive evidence
(3) the total evidence is *inconclusive* (the scale tips in neither direction)

The situation in science, and human knowledge in general, is very similar. In deciding whether to believe a proposition, we presumably weigh the evidence for and against that proposition. As a very simple example, consider the hypothesis that

all swans are white.

First, an example of an irrelevant datum would be any non-swan irrespective of its color. For example, if I examine a pig and ascertain that it is white (or any other color), then this datum is irrelevant to the hypothesis that all swans are white. Finding a white pig, or a pig of any color, provides *no information* one way or the other pertaining to whether all swans are white. Next, an example of positive evidence would be a white swan. Finally, an example of negative evidence would be a non-white swan – for example, a black swan.

The following tables summarizes these examples.

datum	status
a white pig	irrelevant
a white swan	positive
a black swan	negative

[2] The word data is used in two different ways in English, according to the etymological purity (literacy) of the speaker. Technically speaking, 'data' is the plural form of 'datum', and accordingly takes a plural verb form; in other words, 'datum' is technically a *count noun*, like 'citizen', rather than a *mass noun* like 'land'. Both count nouns and mass nouns can be compared and measured, the key difference being whether we use a plural or a singular construction. For example, the U.S. has more land (not lands) than Mexico, and more citizens (not citizen) than Mexico. Alternatively, Mexico has less land, and *fewer* citizens. Notwithstanding the technicality, 'data' is also used colloquially as a mass noun, so that we can say 'less data' rather than 'fewer data'. According to this usage, the implicit metaphysical hypothesis is that data do not ultimately divide into individuals.

The reader has no doubt noticed an asymmetry in the situation. This brings us to the distinction between *conclusive* and *inconclusive* evidence.

evidence E is *conclusive* with respect to hypothesis H $=_{df}$
(1) E logically entails the truth of H; OR
(2) E logically entails the falsity of H

evidence E is *inconclusive* with respect to hypothesis H $=_{df}$
E is not conclusive with respect to H

We can write this symbolically as follows.

E is *conclusive* w.r.t. H $\quad =_{df} \quad E \vdash H$ or $E \vdash \sim H$

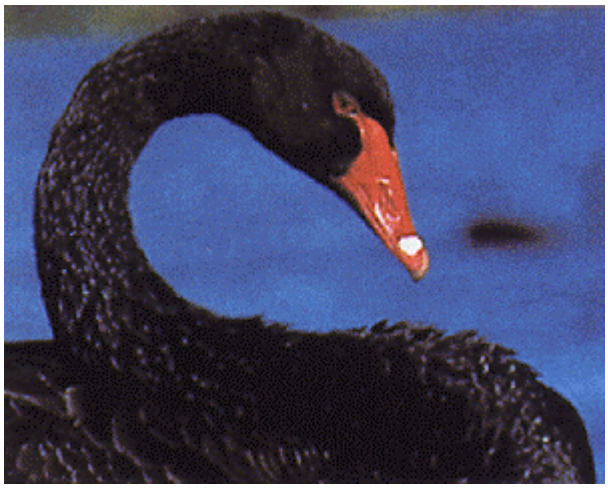
Here, we use the symbol '$\vdash$' ("turnstile") as an abbreviation for 'logically entails'. This gives rise to the subsidiary concepts of conclusive positive evidence and conclusive negative evidence, which are defined as follows.

E is *conclusive positive evidence* for H $\quad =_{df} \quad E \vdash H$
E is *conclusive negative evidence* against H $\quad =_{df} \quad E \vdash \sim H$

For example, finding a black swan provides *conclusive negative evidence* against the hypothesis that all swans are white. This is because the existence of a black swan *logically entails* that not all swans are white. In particular, the following reasoning is valid in the logic of color terms.

this is a black swan;
therefore, this is a swan that is not white;[3]
therefore, there exists at least one swan that is not white;
therefore, **not** all swans are white

In this connection, I offer the following photographic evidence.[4]



[3] This step is not valid in ordinary predicate logic, since the argument form "Bt; therefore ~Wt" is not a valid form. On the other hand, it is a valid argument in the "logic of color terms", which includes the following logical principle: no object is both white and black (all over).
[4] This photo was taken in Perth, Australia [http://www.aaconvperth.org.au/wa.html]. Of course, Australia is home to numerous oddities, including mammals that lay eggs!

Whereas finding a black swan provides conclusive negative evidence against the hypothesis that all swans are white, finding a white swan does not provide *conclusive positive evidence* for this hypothesis. This is because the existence of a white swan does not logically entail that all swans are white. In particular, the following reasoning is *not* valid (the first step is ok, but not the second).

> this is a white swan;
> therefore, there exists at least one white swan;
> therefore, all swans are white.

Returning to our scale metaphor, any conclusive negative datum outweighs *any number* of positive data, and any conclusive positive datum outweighs *any number* of negative data. For example, it does not matter how many white swans you find, a single non-white swan will outweigh them in respect to the hypothesis that all swans are white![5]

In order to obtain conclusive positive evidence for a universal proposition, such as "all swans are white", we need much stronger evidence, which is usually unavailable practically-speaking. The following would be an example. Suppose we somehow manage to observe *every* swan; furthermore, suppose that every observed swan is in fact white. Under these imagined circumstances, we would have conclusive evidence that every swan is white. This is because the following argument is valid.

> every swan is observed;
> every observed swan is white;
> therefore, every swan is white.

The obvious issue is the first premise. How could we ever be warranted in saying that we have observed every swan? This doesn't seem to be scientifically achievable, based on two very general problems – space and time. Concerning space, it is possible that other planets (circling far off stars) have organisms that qualify as swans[6]; the problem is that observing all such organisms would be *practically* impossible. Concerning time, there are two problems – past swans, and future swans. For example, most past swans have left no physical record providing evidence one way or the other about their color. Furthermore, even if we had thousands of records of swans (and especially their feathers), we could never be certain that we have observed all past swans. The problem of future swans is worse. Logically speaking, if not biologically or cosmologically speaking, there are *potentially* infinitely-many future swans. At any given point in our observations, there are swans we have not examined – for the simple reason that they don't exist yet.

We could solve this problem linguistically – by simply declaring that 'swan' means '*current* swan'. The problem then is that our hypothesis lacks *predictive value*, either with respect to future observations of past swans (i.e., their fossils), or with respect to future observations of future swans, and accordingly is considerably less interesting scientifically.

So far, we have seen that, whereas we can find conclusive negative evidence against a universal hypothesis, we cannot find conclusive positive evidence for such an hypothesis. The obverse situation occurs when we consider existential hypotheses, such as the following.

---

[5] If you like, you can describe this as saying that, whereas the weight of inconclusive evidence is finite, the weight of conclusive evidence is "infinite".

[6] It is overwhelmingly improbable that extra-terrestrial swan-like creatures would qualify as swans in the *genetic classification*, according to which w classify organisms according to ancestry, since it is overwhelmingly improbable that these creatures would be ancestrally related to terrestrial swans. Nevertheless, they might nevertheless count as swans in a *phenetic* classification, according which we classify organisms according to their innate traits.

there are organisms that live in volcanoes

If we actually find an organism living in a volcano, this datum constitutes conclusive positive evidence for this hypothesis. On the other hand, if we fail to find such an organism, this does not provide conclusive negative evidence against this hypothesis.

In the absence of conclusive evidence, we must accept a less satisfying position – having varying amounts of positive (or negative) evidence. For example, each white-swan observation provides positive evidence for the hypothesis that all swans are white. Each such observation is said to *confirm* the hypothesis. This is somewhat unfortunate terminology, since it suggests that a single positive instance is *conclusive* to the status of a universal proposition; but it isn't. Indeed, given the potential infinity of relevant data, no finite number of positive instances can be conclusive. Rather, as the number of confirming instances grow, the more confident we become in the given hypothesis.

2. Hempel's Paradox of Confirmation

The above account of confirmation faces a serious logical problem, in the form of Hempel's Paradox.[7] First, in the context of philosophy, a paradox can be generally defined as follows.

a *paradox* is a situation in which:
apparently acceptable/plausible premises
lead by apparently acceptable/plausible reasoning
to an apparently **un**acceptable/**im**plausible conclusion.

A more strict definition (or a special kind of paradox) is given as follows.

a *paradox* is a situation in which:
apparently acceptable/plausible premises
are in logical conflict with each other.

In order to reconstruct Hempel's Paradox of Confirmation, also known as The Raven Paradox, we propose the following three principles.

(p1) a universal proposition "every $\mathbb{F}$ is $\mathbb{G}$" is *confirmed* (to a degree) by any positive instance, where a *positive instance* of "every $\mathbb{F}$ is $\mathbb{G}$" is any *individual* that is both $\mathbb{F}$ and $\mathbb{G}$ [alternatively, a *datum* of the form "this is both $\mathbb{F}$ and $\mathbb{G}$"].
in symbols: $\mathbb{F}t \,\&\, \mathbb{G}t$ confirms $\forall x\{\mathbb{F}x \rightarrow \mathbb{G}x\}$

(p2) a universal proposition "every $\mathbb{F}$ is $\mathbb{G}$" is *not confirmed* (to *any* degree) by any *individual* that is not $\mathbb{F}$ [alternatively, any *datum* of the form "this is not $\mathbb{F}$, and Φ" where Φ is any proposition.]
in symbols: $\sim\mathbb{F}t \,\&\, \Phi$ does not confirm $\forall x\{\mathbb{F}x \rightarrow \mathbb{G}x\}$

(p3) suppose: datum d is logically equivalent to datum d^*
suppose hypothesis H is logically equivalent to hypothesis H^*
then: d confirms H iff d^* confirms H^*

In order to see how these principles are in conflict with one another, let us consider the following hypotheses and data.

[7] Hempel, *Aspects of Scientific Explanation* (1945).

H	=	every raven is black	$\forall x \{ Rx \rightarrow Bx \}$
d	=	t is not a raven, and t is not black	$\sim Rt \ \& \ \sim Bt$
H^*	=	every non-black thing is a non-raven	$\forall x \{ \sim Bx \rightarrow \sim Rx \}$
d^*	=	t is not black, and t is not a raven	$\sim Bt \ \& \ \sim Rt$

We can now deduce a contradiction.

(1)	d does not confirm H	apply (p2), substituting R for $\mathbb{F}$, and $\sim Bt$ for Φ
(2)	d^* confirms H^*	apply (p1), substituting $\sim B$ for $\mathbb{F}$, and $\sim R$ for $\mathbb{G}$
(3)	d is logically equivalent to d^*	sentential logic
(4)	H is logically equivalent to H^*	predicate logic
(5)	d confirms H	2-3 + (p3)
(6)	✕	1,5, sentential logic

The following diagram illustrates the situation.

$\forall x \{ Rx \rightarrow Bx \}$	$\equiv$	$\forall x \{ \sim Bx \rightarrow \sim Rx \}$
$\nvdash$		$\vdash$
$\sim Rt \ \& \ \sim Bt$	$\equiv$	$\sim Bt \ \& \ \sim Rt$

Here, the symbols are read as follows.

$\equiv$	logical equivalence
$\prec$	confirmation
$\not\prec$	non-confirmation.

3. How to Deal with a Paradox

Since every paradox consists of three components –

(1) seemingly acceptable premises
(2) seemingly acceptable reasoning
(3) seemingly unacceptable conclusion

there are basically three approaches to dealing with a paradox.

(1) reject one or more of the premises.
(2) reject the reasoning.
(3) accept the conclusion.

A paradox is like a disease; not only do we have to *diagnose* the underlying problem, we also have to *cure* it, or at least we have to render it more palatable.

1. An Example

By way of illustration of a simpler case, in which both the "diagnosis" and the "cure" are generally well-known, let us review what is oftentimes called 'the paradox of infinity', which we have considered in an earlier chapter.[8] This paradox can be described as the incoherence of the following seemingly plausible principles.

(p1) every object is bigger than all its (proper) parts;
if A (properly) contains B, then A is bigger than B.

(p2) if two sets can be placed in a one-to-one correspondence, then they are equally-big.

(p3) two things are equally-big if and only if neither is bigger than the other.

The problem with these three principles becomes evident when we consider infinite sets – for example:

the set of natural numbers	$\mathbb{N}$	{0, 1, 2, 3, …}
the set of even numbers	$\mathbb{E}$	{0, 2, 4, 6, …}

First of all, it is evident that $\mathbb{E}$ is properly included in $\mathbb{N}$; whereas every member of $\mathbb{E}$ is a member of $\mathbb{N}$, not every member of $\mathbb{N}$ is a member of $\mathbb{E}$. Accordingly, applying (p1) we must conclude that $\mathbb{N}$ is bigger than $\mathbb{E}$; indeed, in some sense, $\mathbb{N}$ is twice as big as $\mathbb{E}$. Next, one can easily establish a one-to-one correspondence between $\mathbb{N}$ and $\mathbb{E}$ – the following being the most obvious one.

$\mathbb{N}$	0	1	2	3	…
	⇳	⇳	⇳	⇳	
$\mathbb{E}$	0	2	4	6	…

Therefore, by (p2) we must conclude that $\mathbb{N}$ and $\mathbb{E}$ are equally-big. Next, by applying (p3) to this conclusion, we obtain that $\mathbb{N}$ is not bigger than $\mathbb{E}$, which contradicts our earlier claim that $\mathbb{N}$ is bigger than $\mathbb{E}$.

In dealing with this paradox, the standard solution is to reject principle (p1). This is the diagnosis. The cure is to re-educate our intuitions about size. In particular, (p1) only *seems* plausible because we base our intuitions on examples of finite objects with finitely-many parts.

4. Back to Ravens

In the Raven Paradox, it is very doubtful that the reasoning is faulty, and the conclusion is an outright contradiction, so accepting the conclusion seems completely out of the question. That means that our only option is to reject one or more of the premises. What is the diagnosis? What is the cure? Here, there is no general consensus among philosophers. Rather, there are a number of different approaches that have been proposed. We discuss a few, and *very* briefly.

[8] "Infinite Sets and Infinite Sizes".

1. Reject (p2)

According to Hempel himself, the conclusion to draw from the above logical exercise is that principle (p2) is incorrect. That is the diagnosis; what is the cure? According to Hempel, although a white swan does not *seem* to be relevant to the hypothesis that all ravens are black, it is! In fact, a white swan provides a *very* small, perhaps even infinitesimal, degree of support for the hypothesis that all ravens are black. Those of us who judge it as irrelevant do so because we regard an infinitesimal confirmation to be no confirmation; this is understandable, since for all practical purposes, infinitesimal *might as well be* none.

Very few philosophers are completely pleased with this solution, so strong is the intuition that a non-raven is *completely* irrelevant to whether every raven is black. For this reason, other approaches have been considered.

2. Adjust (p3) by adjusting our notion of Logical Equivalence.

One alternative approach is to reject principle (p3) – based on the idea that standard logical equivalence is simply too coarse-grained for these purposes. For example, the proposition "all ravens are black" is *about* ravens, whereas the proposition "all non-black things are non-ravens" is *about* non-black things. Accordingly, these two propositions are only equivalent in a very weak sense; they are not equivalent in a strong enough sense for purposes of confirmation. However, this approach provides an acceptable solution only if we can offer an alternative account of strong equivalence that enables us to re-write (p3) so that it is plausible, but does not engender its own paradox.

3. Restrict (p1) and (p2) to Natural Kind Terms

Still another solution adjusts both principles (p1) and (p2), by restricting what counts as an admissible substitution for the letters $\mathbb{F}$ and $\mathbb{G}$. In particular, the proposal is to restrict principles (p1) and (p2) to *natural kind terms*. This serves as a solution, granted we accept the following further premises.

(n1) ‘raven’ is a natural kind term
(n2) ‘black’ is a natural kind term
(n3) ‘non-raven’ is a **not** a natural kind term
(n4) ‘non-black’ is a **not** natural kind term

Then, a non-black non-raven does not confirm "all non-black things are not ravens" because these do not involve natural kind terms.

This accords fairly well with our pre-philosophical intuitions. Most people are even "put off" when asked to consider "goofy" predicates such as ‘non-raven’ and ‘not-black’. On the other hand, what the average English speaker regards as goofy does not seem like an infallible guide to the natural world. For example, the terms of elementary particle physics – e.g., ‘top quark’ ‘charm quark’ and ‘strange quark’ – seem goofy to the outside observer. In spite of their whimsical nature, these terms are nonetheless key components of a highly regarded model of elementary particles.

What we need is a philosophically adequate account of natural kinds that (1) explains natural kinds, and (2) yields (n1)-(n4) as theorems.

4. Adjust (p1) and (p2) to include Background Information

Still another solution adjusts principle (p1) by introducing the notion of "background information". In particular, a datum does not confirm or disconfirm an hypothesis *simpliciter*, but only relative to a prior body of background knowledge. Hypothesis testing does not occur in an epistemological vacuum. In particular, each new observation provides incrementally less support – the first black-raven observation provides more support than the second one, which provides more support than the third one, and so on. This accords with our intuitive sense of the epistemological situation. Consider our exploration of Mars. Suppose we find an extra-terrestrial life form on Mars.[9] This will be newsworthy, to put it mildly. Further suppose that, over the years we discover more and more life forms. Each new life form will be less newsworthy than the previous one, until at some point we will say "ho hum… another life form on Mars". This does not exactly parallel the raven logically, but there is a parallel in terms of expectations. Our first black raven observation has the greatest significance; but with each further black raven datum, we begin to expect that they are all black, so that eventually we no longer find black ravens to be surprising; it becomes a "ho hum" experience.

In order to render precise this intuition, we need to appeal to the notion of probability and information. In particular, we propose that in order for a datum to confirm an hypothesis, the datum must be informative. In order for a datum to be informative, it must have a *prior probability* that is appreciably less than one.[10] This is based on the standard view that information is negative probability – the less likely a proposition, the more information the proposition conveys.

How does this fit with the raven example. Well, if you pick an object at random from the general inventory of objects, it is overwhelmingly probable that you will pick a non-black non-raven (try it!). Accordingly, producing[11] a non-black non-raven provides no information, and accordingly does not confirm *any* hypothesis. On the other hand, producing a black raven provides information, since it is overwhelmingly unlikely that an object picked at random would be a black raven.

5. Another Example

Another example might be useful in undermining our faith in Principle (p1). Suppose that the hypothesis under consideration is:

there are no snakes in Ireland[12]

According to classical and medieval logic, this is a universal proposition, in particular a universal negative proposition. This characterization is supported by modern logic by the fact that such a proposition *may* be written as a universal conditional formula, as follows.

$$\forall x \{ Sx \rightarrow \sim Ix \}$$

According to (p1), this is confirmed (to a degree) by any positive instance. So, we can confirm it by finding a snake that is not in Ireland; furthermore, we can pile up supporting evidence by finding lots

[9] Most likely, it will be an extinct one-celled organism, but even that will be amazing.
[10] By appreciably less than one, I mean simply finitely-less than one, as opposed to infinitesimally-less than one.
[11] Here, 'produce' simply means 'bring forth, or exhibit'; it does not mean 'create' or 'manufacture'.
[12] We understand an Irish snake to be a snake that is *native* to Ireland, not a pet or a zoo specimen, even one that has escaped. The well-known story is that Saint Patrick (a Roman nobleman – i.e., patrician – hence his name) rid Ireland of its snakes. A more plausible (evolutionary) account is that the water between the main island of Britain and Ireland is too cold for snakes to cross.

and lots of snakes that aren't in Ireland. But this is surely absurd! Finding any number of snakes not in Ireland is completely irrelevant to whether there are snakes in Ireland. Indeed, as I encounter more and more snakes, especially if they are scattered throughout the world (but not in Ireland), perhaps I *should* begin to suspect that the hypothesis is false. I start to wonder "why not Ireland too?"

This intuition is moreover supported by the information-theoretic account of confirmation, according to which a non-Irish snake does not confirm *any* hypothesis. The reason for this is that we believe already that an overwhelming majority of things are not in Ireland; so, finding a snake not in Ireland is simply *not newsworthy*.

What if we look at Irish things and confirm they are not snakes. Once again, finding a non-snake Irish inhabitant is uninformative, because the *a priori* expectation is that nearly anything you find in Ireland will be a non-snake. So *merely* producing an Irish non-snake does not constitute evidence that there are no snakes in Ireland.

So, what *would* count as evidence for the hypothesis that there are no snakes in Ireland? Well, it is first of all obvious that we have to go to Ireland; no searches for snakes on other land masses will do. Second, once we are in Ireland, we have to make a *serious effort* to find snakes there. If we search for snakes in Ireland, and we fail to find any, this constitutes good evidence that there aren't any snakes in Ireland. How good the evidence is depends upon how good the search is. For example, if we confine our search to pubs, this would be a rather poor search (for snakes at least!), and the evidence would be correspondingly weak.

The philosophical problem is that the concept of a "serious search" does not admit a purely-logical description. This strongly suggests that confirmation is not a purely-logical concept.