Induction: a formal perspective

Uwe Saint-Mont, Nordhausen University of Applied Sciences[*]

December 2, 2019

Abstract. The aim of this contribution is to provide a rather general answer to Hume's problem. To this end, induction is treated within a straightforward formal paradigm, i.e., several connected levels of abstraction.

Within this setting, many concrete models are discussed. On the one hand, models from mathematics, statistics and information science demonstrate how induction might succeed. On the other hand, standard examples from philosophy highlight fundamental difficulties.

Thus it transpires that the difference between unbounded and bounded inductive steps is crucial: While unbounded leaps of faith are never justified, there may well be reasonable bounded inductive steps.

In this endeavour, the twin concepts of information and probability prove to be indispensable, pinning down the crucial arguments, and, at times, reducing them to calculations.

Essentially, a precise study of boundedness settles Goodman's challenge. Hume's more profound claim of seemingly inevitable circularity is answered by obviously non-circular hierarchical structures.

Keywords

Hume's Problem; Induction; Inference

[*] Prof. U. Saint-Mont, Fachbereich Wirtschafts- und Sozialwissenschaften, Hochschule Nordhausen, Weinberghof 4, D-99734 Nordhausen; saint-mont@hs-nordhausen.de; ORCID: 0000-0001-6801-3658.

Contents

1 Introduction

A problem is difficult if it takes a long time to solve it; it is important if a lot of crucial results hinge on it. In the case of induction, philosophy does not seem to have made much progress since Hume's time: Induction is still the glory of science and the scandal of philosophy (Broad (1952), p. 143), or as Whitehead (1926), p. 35, put it: "The theory of induction is the despair of philosophy - and yet all our activities are based upon it." Since a crucial feature of science is general theories based on specific data, i.e., some kind of induction, Hume's problem seems to be both: difficult and important.

Let us first state the issue in more detail. Traditionally, many dictionaries define inductive reasoning as the derivation of general principles/laws from particular/individual instances. For example, according to the Encylopedia Britannica (2018), induction is the "method of reasoning from a part to a whole, from particulars to generals, or from the individual to the universal." However, nowadays, philosophers rather couch the question in 'degrees of support.' Given valid premises, a deductive argument preserves truth, i.e., its conclusion is also valid. An inductive argument is weaker, since such an argument transfers true premises into some degree of support for the argument's conclusion. The truth of the premises provides (more or less) good reason to believe the conclusion to be true.

Although these lines of approach could seem rather different, they are indeed very similar if not identical: Strictly deductive arguments, preserving truth, can only be found in logic and mathematics. The core of these sciences is the method of proof which always proceeds (in a certain sense, later described more explicitly) from the more general to the less general. Given a number of assumptions, an axiom system, say, any valid theorem has to be derived from them. That is, given the axioms, a finite number of logically sound steps imply a theorem. In this sense, a theorem is always more specific than the whole set of axioms; its content is more restricted than the complete realm defined by the axioms: Euklid's axioms define a whole geometry, whereas Phythagoras' theorem just deals with a particular kind of triangle.

Induction fits perfectly well into this picture: Since there are always several ways to generalize a given set of data, "there is no way that leads with necessity from the specific to the general" (Popper). In other words, one cannot prove the move from the more specific to the less specific. A theorem that holds for rectangles need not hold for arbitrary four-sided figures. Deduction is possible if and only if we go from general to specific. When moving from general to specific one may thus try to strengthen a non-conclusive argument until it becomes a proof. The reverse to this is, however, impossible. Strictly non-deductive arguments, those that cannot be 'fixed' in principle, are those which universalise some statement.

2 Hume's problem

2.1 Verbal exposition

Gauch (2012), pp. 168 gives a concise modern exposition of the issue and its importance:

> (i) Any verdict on the legitimacy of induction must result from deductive or inductive arguments, because those are the only kinds of reasoning.

(ii) A verdict on induction cannot be reached deductively. No inference from the observed to the unobserved is deductive, specifically because nothing in deductive logic can ensure that the course of nature will not change.

(iii) A verdict cannot be reached inductively. Any appeal to the past successes of inductive logic, such as that bread has continued to be nutritious and that the sun has continued to rise day after day, is but worthless circular reasoning when applied to induction's future fortunes

Therefore, because deduction and induction are the only options, and because neither can reach a verdict on induction, the conclusion follows that there is no rational justification for induction.

Notice that this claim goes much further than some question of validity: Of course, an inductive step is never sure (may be invalid); Hume, however, disputes that inductive conclusions, i.e., the very method of generalizing, can be justified at all. Reichenbach (1956) forcefully pointed out why this result is so devastating: Without a rational justification for induction, empiricist philosophy in general and science in particular, hang in the air. However, worse still, if Hume is right, such a justification quite simply does not exist. If empiricist philosophers and scientists only admit empirical experiences and rational thinking, they have to contradict themselves since, at the very beginning of their endeavour, they need to subscribe to a transcendental reason (i.e., an argument neither empirical nor rational). Thus, this way to proceed is, in a very deep sense, irrational.

Consistently, Hacking (2001), p. 190, writes: "Hume's problem is not out of date [...] Analytic philosophers still drive themselves up the wall (to put it mildly) when they think about it seriously." For Godfrey-Smith (2003), p. 39, it is "The mother of all problems." Moreover, quite obviously, there are two major ways to respond to Hume:

(i) Acceptance of Hume's conclusion. This seems to have been philosophy's mainstream reaction, at least in recent decades, resulting in fundamental doubt. Consequently, there is now a strong tradition questioning induction, science, the Enlightenment, and perhaps even the modern era.

(ii) Challenging Hume's conclusion and providing a more constructive answer. This seems to be the typical way scientists respond to the problem. Authors within this tradition often concede that many particular inductive steps are justified. However, since all direct attempts to solve the riddle seem to have failed, there is hardly any general justification. Tukey (1961) is quite an exception: "Statistics is a broad field, whether or not you define it as 'The science, the art, the philosophy, and the techniques of making inferences from the particular to the general'."

Since the basic viewpoints are so different, it should come as no surprise that, unfortunately, clashes are the rule and not the exception. For example, when philosophers Popper und Miller (1983) tried to do away with induction once and for all, physicist Jaynes (2003), p. 699, responded: "Written for scientists, this is like trying to prove the impossibility of heavier-than-air flight to an assembly of professional airline pilots."

2.2 Formal treatment

A straightforward paradigmatic model consists of two levels of abstraction and the operations of deduction and induction connecting them. That is, in the following illustration the more general tier on top contains more information than its more specific counterpart at the bottom. Moving downwards, deduction skips some of the information. Moving upwards, induction leaps from 'less to more:'

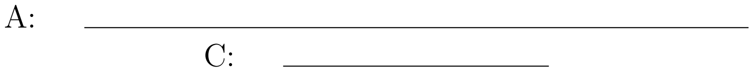


Fig. 1 Basic model with two tiers. A (more general), and C (less general).

Here is another interpretation: The notion of generality is crucial in (and for) the world of mathematics. Conditions operate 'top down', each of them restricting some situation further. The 'bottom up' way is constructive, with sets of objects becoming larger and larger. Since almost all of contemporary mathematics is couched in terms of set theory, the most straightforward way to encode generality in the universe of sets is by means of the subset relation. Given a certain set B, any subset C is less general, and any superset A is more general ($C \subseteq B \subseteq A$).

Although this model is hardly more than a reformulation of the original problem, it has the advantage of delimiting the situation. Instead of pondering a vague inductive leap of faith, it introduces two rather well-defined *layers* and the *gap* between them. Consistently, one is led to the idea of a distance $d(A, C)$ between the layers, and it is straightforward to distinguish three major cases:

Basic classification

(i) $d(A, C) = 0$, i.e., A and C coincide

(ii) $d(A, C) \leq b$, where b is a finite number, i.e., the distance is bounded

(iii) $d(A, C) = \infty$, i.e., the distance is unbounded

Obviously, there is no need for induction in the first case. Mathematically speaking, if A implies C, and vice versa, A and C are equivalent. The second case motivates well-structured inductive leaps - there could be justifiable 'small' inductive steps. However, given an infinite distance, a leap of faith from C to A never seems to be well-grounded.

Notice that, although the latter classification looks rather trivial, the division it proposes is a straightforward consequence of our basic model which hardly deviates from the received problem. In other words: Given the classical problem of induction, the above division is almost inevitable.

3 The many faces of induction

3.1 Philosophy

Groarke (2009), pp. 80, 87, 79 (my emphasis) restates the basic model in rather philosophical terminology: "We have, then, two metaphysics. On the Aristotelian, substance, understood as the true nature of existence of things, is open to view. The world can be observed. On the empiricist, it lies underneath perception; the true nature of reality lies within an invisible substratum forever closed to human penetration. . .To place substance, ultimate existence, *outside the limits* of human cognition, is to leave us enough mental room to doubt anything. . .It is the *remoteness* of this ultimate metaphysical reality that undermines induction."

Apart from this rather roundabout treatment, particular situations have been studied in much detail:

Eliminative induction

A classical approach, preceding Hume, is *eliminative induction.* Given a number of hypotheses on the (more) abstract layer A and a number of observations on the (more) concrete layer C, the observations help to eliminate hypotheses. In detective stories, with a finite number of suspects (hypotheses), this works fine. The same applies to an experimentum crucis that collects data in order to decide between just *two* rival hypotheses.

However, the real problem seems to be unboundedness. For example, if one observation is able to delete k hypotheses, an infinite number of hypotheses will remain if there is an infinite collection of hypotheses but just a finite number of observations. That is one of the main reasons why string theories in modern theoretical physics are notorious: On the one hand, due to their huge number of parameters, there is an abundance of different theories. On the other hand, there are hardly any (no?) observations that effectively eliminate most of these theories (Woit 2006).

More generally speaking: If the information in the observations suffices to narrow down the number of hypotheses to a single one, eliminative induction works. However, since hypotheses have a surplus meaning, this could be the exception rather than the rule.

Enumerative induction

Perhaps the most prominent example of a non-convincing inductive argument is Bacon's *enumerative induction.* That is, do a finite number of observations $x_1, \ldots, x_n$ suffice to support a general law like "all swans are white"? A similar question is if/when it is reasonable to proceed to the limit $\lim x_i = x$.

Given as little as a finite sequence, the infinite limit is, of course, arbitrary. Therefore the concept of mathematical convergence is defined the other way around: Given an infinite sequence, any finite number of elements is irrelevant (e.g., the first n). A sequence $(x_i)_{i \in I\!N}$ converges towards x if 'almost all' x_i (all except a finite set of elements) lie within an arbitrary small neighborhood of x. That is, given some distance measure $d(x, y)$ and any $\epsilon > 0$, then for almost all i, $d(x_i, x) < \epsilon$. This is equivalent to saying that for every $\epsilon > 0$ there is an $n(\epsilon)$ such that all x_i with $i \geq n$ do not differ more than ϵ from x, that is, for all $i \geq n$ we have $d(x_i, x) < \epsilon$.

Our interpretation of this situation amounts to saying that any finite sequence contains a very limited amount of information. If it is a binary sequence of length n, exactly n yes-no questions have to be answered in order to obtain a particular sequence $x_1, \ldots, x_n$. In the case of an arbitrary infinite binary sequence $x_1, x_2, x_3, \ldots$ one has to answer an infinite number of such questions. In other words, since the gap between the two situations is not bounded, it cannot be bridged. In this situation, Hume is right when he claims that "one instance is not better than none", and that "a hundred or a thousand instances are ... no better than one" (cf. Stove (1986), pp. 39-40).

Further assumptions are needed, either restricting the class of infinite sequences or strengthening the finite sequence considerably. Either way, the gap between A and C becomes smaller, and one may hope to get a reasonable result if the additional assumptions render the distance finite. For a thorough discussion see section 5.1.

Frequentist probability

Trying to define probability in terms of a limit of empirical frequencies is a typical example of how *not* to treat the problem. Empirical observations - of course, always a finite number - may have a 'practical limit,' i.e., they may stabilise quickly. However, that is not a limit in the mathematical sense requiring an infinite number of (idealized) observations.[1] Trying to use the empirical observation of 'stabilisation' as a definition of probability (Reichenbach 1938, 1949), inevitably needs to evoke infinite sequences, a mathematical idealization.

Thus the frequentist approach easily confounds the theoretical notion of probability (a mathematical concept) with limits of observed frequencies (empirical data). In the same vein highly precise measurements of the diameters and the perimeters of a million circles may give a good approximation of the number π; nevertheless, physics is not able to prove a single mathematical fact about π. Instead, mathematics must define a circle as a certain *relation of ideas*, and also needs to 'toss a coin' in a theoretical framework. A contemporary and logically sound treatment is given in section 5.1.

3.2 Computer Sciences

Quite distinctive inductive problems can be found in this area. They range from the smallest inductive step perceivable to malign situations:

An elementary model

The basic unit of information is the Bit. This logical unit may assume two distinct values (typically named 0 and 1). Either the state of the Bit B is known or set to a certain value, (e.g., $B = 1$), or the state of Bit B is not known or has not been determined, that is, B may be 0 or 1. The elegant notation used for the latter case is $B =?$, the question mark being called a "wildcard."

In the first case, there is no degree of freedom: We have particular data. In the second case, there is exactly one (elementary) degree of freedom. Moving from the general case with one degree of freedom to the special case with no degree of freedom is simple: Just answer the following yes-no question: "Is B equal to 1, yes or no?" Moreover, given a number of Bits, more or less general situations can be distinguished in an extraordinarily simple way: One just counts the number of yes-no questions that need to be answered (and that's exactly what almost any introduction to information theory does), or, equivalently, the number of degrees of freedom lost. Thus, beside its elegance and generality, the major advantage of this approach is the fact that everything is finite, allowing interesting questions to be answered in a definite way.

[1] The essential point of the mathematical definition is the behaviour of almost all members of some sequence (i.e., all but a finite number). Therefore any number of empirical observations is not able to bridge the gap between strictly finite sequences and the realm of infinite sequences.

Consider the following example:

most general:	??????
general:	101???
specific:	1010?0
most specific:	101000

Reading these lines from top to bottom, one observes that whenever a wildcard (?) is replaced by a particular number in a certain column, this number does not change further down. Thus there is a trivial kind of deduction: Given a particular number on a certain tier implies (without any doubt) that this number also shows up on all lower tiers.

Vice versa, reading the table upwards can be understood in an inductive way. In the most elementary case, a step further up is tantamount to replacing a single concrete digit by a wildcard (e.g., $B = 1$ by $B =?$). This reveals a non-trivial fork-like structure: The particular value of B splits into two possibilities. Looking at the whole sequence indicates why quite a few have wondered about induction. Indeed, it is Janus-faced: On the one hand there are digits that do not change, and this partial stability (i.e., stability in some digits) resembles deduction. On the other hand, if a concrete digit is replaced by the wildcard, the move upwards is perfectly non-deductive. The character '?' signifies just that: We lose the information in this digit, i.e., we cannot conclude with any degree of certainty which number is the correct one.

A refined model: distributions

Most authors who consider Hume's problem discuss probability theory at some point. Here, it is straightforward to evoke the notion of a random variable having a certain distribution. For example, flipping a coin is tantamount to a random variable X assuming the values $X = 1$ (Heads) and $X = 0$ (Tails) with probabilities p and $1 - p$, respectively.[2] If one knows which of the two values is the case one speaks of a certain realization of the random variable. If not, the obtainable values plus their probabilities form the distribution of the random variable.

That is indeed very similar to the above model. Going from the distribution to the realization corresponds to a step downwards, and moving upwards depicts very nicely the bifurcation one encounters. Since the degrees of freedom rise when moving upwards, but also since we have augmented the above model (replacing a wildcard by a certain distribution) we have to determine the value of p. Owing to symmetry that is, because 0 and 1 have exactly the same status, $p = 1/2$ is the most natural choice.

A reversed view

A more typical way to read this situation is, however, just the reverse. Since it is possible to move without doubt from a more specific (informative, precisely described) situation to a less specific situation, we know that if 101000 is true, so must be 101???. It is no problem whatsoever to skip or blur information, e.g., to move from a precise quantitative statement to a roundabout qualitative one. And that's exactly what happens here upon moving upwards. Deduction means to lose some information or at best to keep the information content unchanged. The content of a statement becomes less, or, in the case of tautology, no information gets lost. Upon moving up, we know less

[2] In statistical jargon, X has a Bernoulli distribution $B(p)$ with parameter p.

and less about the specific pattern being the case. In this view, every ? stands for 'information unavailable.' The more question marks, the less we know, and thus our knowledge becomes meagre upon moving up.

This also means that the other direction is not trivial, is more difficult and interesting: In order to move further down, one has to generate information. Thus we have to ask yes-no questions, and their answers provide precisely the information needed. In this view, an elementary move downwards replaces the single sign ? by one of the concrete numbers 0 and 1. That's also a kind of bifurcation, and an inductive step, since the amount of information in the pattern increases. Thus the most specific pattern right at the bottom contains a maximum of information, and it also takes a maximum number of inductive steps to get there. In a picture, we get a trapezoid with the shorter side at the top:

??

0? 1?

00 01 10 11

Synthesis

The models considered in this section demonstrate that moving from the general to the particular and back need *not* involve a major drawback. Rather, the framework just elaborated exemplifies *minimal* inductive steps. Another insight is that such an inductive step can be treated elegantly with the help of probability theory (a single number or sign may be replaced by a less informative two-point distribution), and that information and probability are closely related.

However, it turns out that, depending on how one defines more or less specific, induction comes in when moving up or down. Upwards, a specific number (e.g., 1) is replaced by a less focused distribution (e.g., $B(1/2)$). Downwards, a single sign (?) is split into two numbers (that could also be augmented to a probability distribution, although we have not done so explicitly). Deduction can also be understood in two ways: The general pattern (e.g., 101???) constitutes a boundary condition for some more concrete sequence further down, therefore the first three digits of such an observation must be 101. Conversely, if 101000 is the case, so must be the pattern ?0?0?0.

Thus, although important, the notion of generality seems to be less crucial than the concept of *information.* Losing information is easy, straightforward, and may even be done algorithmically. Therefore, such a step should be associated with the adjective *deductive.* In particular, it preserves truth. Moving "from less to more" (Groarke (2009), p. 37), acquiring information, or increasing precision is much more difficult, and cannot be done automatically. Thus this direction should be called *inductive.* Combining both directions, a trapezoid, a funnel or a tree-like structure may serve as standard formal representatives for deductive vs. inductive moves (see Fig. 1 but also Section 4.1). Note, however, that there are many possible ways to skip or to add information. Thus, in general, neither an inductive nor a deductive step is unique.

Information theory

Owing to the finite nature of the model(s) just considered, they can be extended to a complete formal theory of induction. In this abstract view, anything - in particular hypotheses, models, data and programs - is just series of zeros and ones. Moreover, they may all be manipulated with the help of computers (Turing machines). A universal computer is able to calculate anything that is computable.

Within this framework, *deduction* of data $\mathbf{x}$ means to feed a computer with a program $\mathbf{p}$, automatically leading to the output $\mathbf{x}$. *Induction* or data *compression* is the reverse: Given output $\mathbf{x}$, find a program $\mathbf{p}$ that produces $\mathbf{x}$. As is to be expected, there is a fundamental asymmetry here: Proceeding from input $\mathbf{p}$ to output $\mathbf{x}$ is straightforward. However, given $\mathbf{x}$ there is no automatic or algorithmic way to find a non-trivial shorter program $\mathbf{p}$, let alone $\mathbf{p}^*$, the smallest such program. Although the content of $\mathbf{p}$ is the same as that of $\mathbf{x}$, there is more redundancy in $\mathbf{x}$, blocking the way back to $\mathbf{p}$ effectively.

However, fundamental doubt has not succeeded here; au contraire, Solomonoff (1964) provided a general, sound answer to the problem of induction. His basic concept is Kolmogorov (algorithmic) compexity $K(\mathbf{x})$, i.e., the length of the shortest prefix-free program p^* delivering output $\mathbf{x}$. In a sense, this theory is just a mathematically refined (and thus logically sound!) version of Occam's razor: "Select the simplest hypothesis compatible with the observed values" (Kemeny (1953), p. 397).[3]

It should be noted that the term *prefix-free*, i.e, "no program is a proper prefix of another programm" (Li and Vitányi (2008), p. 199), is crucial, since one thus avoids circularity. More precisely: If programs are allowed to be prefixes of other programs, programs can be nested into each other which leads to divergent (unbounded) series. Many instructive examples can be found in Li and Vitányi (2008), pp. 197-199.

3.3 Statistics

The paradigm of (standard) statistics quite explicitly consists of two tiers: C - empirical observations or quite simply 'data' on the one hand, and A - some general 'population', for instance, a family of probability distributions (potential 'laws' or 'hypotheses') on the other hand.

Williams' example

Perhaps Williams (1947) was the first philosopher who employed statistics to give a constructive answer to Hume's problem. Here is his argument in brief (p. 97):

> Given a fair sized sample, then, from any population, with no further material information, we know logically that it very probably is one of those which match the population, and hence that very probably the population has a composition similar to that which we discern in the sample. This is the logical justification of induction.

In modern terminology, one would say that most (large enough) samples are typical for the population from whence they come. That is, properties of the sample are close to corresponding properties of the population. In a mathematically precise sense, the distance between sample and population is small. Now, being similar is a symmetric concept: A property of the population can be found (approximately) in the sample and vice versa, i.e., given a sample (due to combinatorial reasons most likely a representative one), it is to be expected that the corresponding value in the population does not differ too much from the sample's estimate. In other words: If the distance between population and sample is small, so must be the distance between sample and population, and this is true in a mathematically precise sense.

[3] For details see the original work and the contemporary and thorough treatments in Cover and Thomas (2006), Li and Vitányi (2008).

Asymptotic statistics

Many philosophers focused on details of Williams' example and questioned the validity of his result (for a review see Stove (1986), Campbell (2001), Campbell and Franklin (2004)). Yet for mathematicians, Williams' rather informal reasoning is sound and can be extended considerably: The very core of asymptotic mathematical statistics and information theory consists in the successful comparison of (large) samples and populations.

Informally speaking, the gap between sample and population isn't merely bounded, rather, it is quite narrow from the beginning. Thus, if the size of the sample is increased, it can typically be reduced to nothing. For example, given a finite population and sampling without replacement, having drawn all members of the population, the sample is exactly equal to the population. But infinite populations are also nothing to be afraid of. For example, if $\hat{\theta} = \theta(x_1, \ldots, x_n)$ is a mathematically reasonable estimator of θ, we have $\hat{\theta} \to \theta$ in the sense that the probability $p(|\hat{\theta} - \theta| < \varepsilon)$ goes to zero for every $\varepsilon > 0$.

Much more generally, there are vast formal theories of hypothesis testing, parameter estimation and model identification. Williams' very specific example works because of the law of large numbers (LLN) which guarantees that (most) sample estimators are consistent, i.e., they converge toward their population parameters in the probabilistic sense just explained. In particular, the relative frequency of red balls in the samples considered by Williams approaches the proportion of red balls in the urn. That's trivial for a finite population and selection without replacement (when you have drawn all balls from the urn you know the proportion), however convergence is guaranteed (much) more generally. One of the most important results of this kind is the main theorem of statistics, i.e., that the empirical distribution function $\hat{F} = F(x_1, \ldots, x_n)$ approximates the 'true', i.e., the population's distribution function F in a strong sense. Convergence is also robust, i.e., it still holds if the seemingly crucial assumption of independence is violated (there are strong convergence results for general stochastic processes, and in certain cases the assumption may even be dropped, see Fazekas and Klesov (2000)). Moreover, the rate of convergence is very fast (see the literature building on Baum et al. (1962)). Extending the orthodox sample-population model to a Bayesian framework (with prior, sample and posterior) also does not change much, since strong convergence theorems exist there too (cf. Walker (2003, 2004)).

In a nutshell, it is difficult to imagine stronger rational foundations for an inductive claim: there is a bounded gap that can be leapt over by lifting the lower level. Within the paradigm of statistics, this is almost tantamount to collecting more observations. Statistics' basic theorems then say that, given mild conditions, samples approximate their populations, the larger n the better (in particular there are the LLN, the main theorem of statistics, and the central limit theorem). Fisher (1935/1966), p. 4, concluded optimistically:

> We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression. . . . The mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference.

Could the gap be bridged by lowering the upper level? Just recently, Rissanen (2007) did so. In essence, his idea is that data contains a limited amount of information. With respect to a family of hypotheses this means that, given a fixed data sample, only a certain number of hypotheses can be reasonably distinguished. Thus he introduces the notion of *optimal distinguishability* which is the number of (equivalence classes of) hypotheses that can be reasonable distinguished: too many such classes and the data do not allow for a decision between two adjoint (classes of) hypotheses with high enough probability; too few equivalence classes of hypotheses means wasting information available in the data.

Widening the gap

When is it difficult to proceed from sample to population, or, more crudely put, from n (large sample) to $n+1$ (the whole population)? Here is one of these cases: Suppose there is a large but finite population consisting of the numbers $x_1 \ldots, x_{n+1}$. Let x_{n+1} be really large (10^{100}, say), and all other x_i tiny (e.g., $|x_i| < \epsilon$, with ϵ close to zero, $1 \leq i \leq n$). The population parameter of interest is $\theta = \sum_{i=1}^{n+1} x_i$. Unfortunately, most rather small samples of size k do not contain x_{n+1}, and thus almost nothing can be said about θ. Even if $k = 0,9 \cdot (n+1)$, about 10% of these samples still do not contain x_{n+1}, and we know almost nothing about θ. In the most vicious case a nasty mechanism picks $x_1, \ldots, x_n$, excluding x_{n+1} from the sample. Although all but one observation are in the sample, still, hardly anything can be said about θ since $\sum_{i=1}^{n} x_i$ may still be close to zero.

Challenging theoretical examples have in common that they *withhold* relevant information about the population as long as possible. Thus even large samples contain little information about the population. In the worst case, a sample of size n does not say anything about a population of size $n+1$. In the example just discussed, $x_1, \ldots, x_n$ has *nothing* to say about the parameter $\theta' = \max(x_1, \ldots, x_{n+1})$ of the whole population. However, if the sample is selected at random, simple combinatoric arguments guarantee that convergence is almost exponentially fast.

Rapid and robust convergence of sample estimators toward their population parameters makes it difficult to cheat or to sustain principled doubt. It needs an intrinsically difficult situation or an unfair 'demonic' selection procedure (Indurkhya 1990) to obtain a systematic bias, rather than to just slow down convergence. Therefore it is no coincidence that other classes of 'unpleasant examples' emphasize intricate dependencies among the observations (e.g., non-random, biased samples), single observations having a large impact (e.g., the contribution of the richest household to the income of a village), or both (e.g., see Érdi (2008), chapter 9.3). That's why, in practice, earthquake prediction is much more difficult than foreseeing the colour of the next raven.

Despite these shortcomings, statistics and information theory both teach that - typically - induction is rationally justified. Although their fundamental concepts of information and probability can often be used interchangeably, it should be mentioned that a major technical advantage of information over probability (and other related concepts) is that information is non-negative and additive (Kullback 1959). Thus information increases monotonically. In the classical 'nice' cases, it accrues *steadily* with every observation. So, given a large enough sample k, much can be said about the population (more precisely, the information I in or represented by the population), since the difference $I - I(k)$ decreases gradually.

4 Extensions

4.1 The multi-tier model

Qualitatively speaking, a bounded distance between A and C is good-natured, and an unbounded distance is not. However, the smaller $d(A, C)$, the better. In other words, a small inductive step is more convincing than a giant leap of faith, and lacking a specific context, an inductive conclusion seems to be justified if $d(A, C)$ is sufficiently small. That is, if the inductive leap is smaller than some threshold t, or when it can be made arbitrarily small in principle.

If the upper layer represents some general law and the lower layer represents data (concrete observations) all that is needed to 'reduce the gap' is to add assumptions (essentially lowering the upper layer, making the law less general) or to add observations (lifting the lower layer upwards, extending the empirical basis). More generally, given (nested) sets of conditions or objects, the corresponding visual display is a "funnel", with the most general superset at the top, and the most specific subset at the bottom:

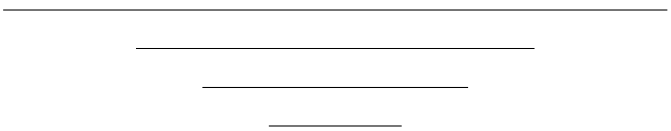


Fig. 2 Several tiers of abstraction (the funnel)

Suppose - corresponding to the initial inductive problem - that the top and the bottom tiers are fixed. Then every new layer in-between makes the inductive gap(s) that have to be bridged smaller, since each such tier replaces a large step by two smaller ones. Iterating this process may turn a gigantic and thus extremely implausible leap into a staircase of tiny steps upwards, making a high mountain accessible, even for a critical mind that only accepts tiny inductive moves. For example, the creation of complex life forms in a single step is nothing short of a miracle. Yet modern science has demystified this issue, since it is able to fill in the details, i.e., it can explain the involved evolutionary development from physical principles via chemical reactions to biological designs - step by step.

Black (1958) went farther: Inductive investigation also means that some line of argument may come in handy elsewhere. That is, some inductive chain of arguments may be supported by another, different, one. In union there is strength, i.e., funnel F may borrow strength from funnel G (and vice versa), such that a conclusion resting on several lines of investigation - pillars - may be further reaching or better supported:

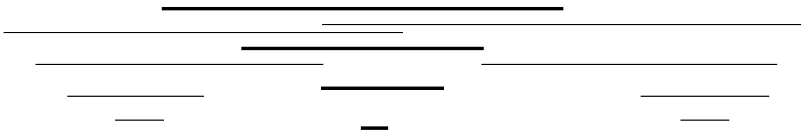


Fig. 3 Several funnels

That is, Fig. 2 is multiplied until the funnels begin to support each other on higher levels of abstraction. A classical example is Perrin (1990) on the various methods of demonstrating the existence of atoms: Since each line of investigation points toward the existence of atoms, it is difficult to escape the conclusion that atoms indeed exist. If they do, one has found the 'overlying' reason for all particular phenomena observed.

4.2 Convergence

Inserting further and further layers in between, an inductive gap can be made - at least in principle - arbitrarily small. If the number of such layers goes to infinity, this brings up the charge of infinite regress, i.e., an endless series of arguments based on each other. However, calculus teaches (or Bolzano and Cauchy if you like) that a bounded series possesses a convergent subseries, and that a bounded monotone series converges. In the present context this means that if the initial inductive gap is bounded, additional assumptions that have to be evoked in order to narrow the inductive steps necessarily tend to become *weaker*. So even if an infinite number of assumptions were needed, most of them would have minuscule consequences.

This solution is very similar to mathematics' answer to Zenon's tale of Achill and the turtle: Suppose Achill starts the race in the origin ($x_0 = 0$), and the turtle at a point $x_1 > 0$. After a certain amount of time, Achill reaches x_1, but the turtle has moved on to point $x_2 > x_1$. Thus Achill needs some time to run to x_2. However, meanwhile, the turtle could move on to $x_3 > x_2$. Since there always seems to be a positive distance between Achill and the turtle ($x_{i+1} - x_i > 0$), a verbal argument will typically conclude that Achill will never reach or overtake the turtle. Of course, practice teaches otherwise, but it took several hundred years and some mathematical subtlety to find a theoretically satisfying answer. Dealing with Hume's problem, Rescher (1980), pp. 208-209 also uses an iterative sequence of inductive arguments. However, although he mentions Achill and the tortoise explicitly, he fails to realize that a precise notion of convergence has dissolved Zenon's paradox.

Note that in the case of an unbounded gap that is completely different: There may be very large (and thus unconvincing) leaps of faith, and chains of assumptions becoming ever stronger. In particular, it is not convincing to justify a concrete inductive step by a weak inductive rule that is vindicated by a stronger inductive law, etc. Consistently, an ultimate, i.e. very strong inductive principle, is the least pervasive, for instance "there is a general rule that 'protects' (some, many, most) particular inductive steps" or "inductive leaps are always justified."

4.3 The general inductive principle

Nevertheless, if, typically, or at least very often, generalizations are successful, inductive thinking (looking for a rule for those many examples) will almost inevitably lead to the idea that there could be some general principle, justifying particular inductive steps:

> There are plenty of past examples of people making inductions. And when they have made inductions, their conclusions have indeed turned out true. So we have every reason to hold that, in general, inductive inferences yield truths (Papineau (1992), p. 14); that is, *it is reasonable to believe that* induction works well in general (and is thus an appropriate mode of reasoning).[4]

At this point it is extremely important to distinguish between concrete lines of inductive reasoning on the one hand, and induction in general on the other. As long as there is a funnel-like structure which can always be displayed a posteriori in the case

[4] A conclusion attributed to Braithwaite (1953) by Rescher (1980), p. 210, his emphasis.

of a successful inductive step, there is no fundamental problem. Generalizing a certain statement with respect to some dimension, giving up a symmetry or subtracting a boundary condition is acceptable, as long as the more abstract situation remains well-defined.[5] The same holds with the improvement of a certain inductive method which is elaborated in Rescher (1980): Guessing an unknown quantity with the help of Reichenbach's straight rule may serve as a starting point for the development of more sophisticated estimation procedures, based on a more comprehensive understanding of the situation. Some of these 'specific inductions' will be successful, some will fail.

But the story is quite different for induction in general! Within a well-defined bounded situation, it is possible to pin down, and thus justify, the move from the (more) specific to the (more) general. However, in total generality, without any assumptions, the endpoints of an inductive step are missing. Beyond any concrete model, the upper and the lower tier, defining a concrete inductive leap, are missing, and one cannot expect some inductive step to succeed. For a rationally thinking person, there is no transcendental reason that a priori protects abstraction (i.e., the very act of generalizing). The essence of induction is to extend or to go beyond some information basis. This can be done in numerous ways and with objects of any kind (sets, statements, properties, etc.). The vicious point about this kind of reasoning is that the straightforward, inductively generated expectation that there should be a general inductive principle overarching all specific generalizations is an inductive leap that fails. It fails since without boundary conditions - any restriction at all - we find ourselves in the unbounded case, and there is no such thing as a well-defined funnel there.[6]

Following this train of thought, Hume's paradox arises since we confuse a well-defined, restricted situation (line 1 of the next table) with principal doubt, typically accompanying an unrestricted framework (or no framework at all, line 2). On the one hand Hume asks us to think of a simple situation of everyday life (the sun rising every morning), a scene embedded in a highly regular scenario. However, if we come up with a reasonable, concrete model for this situation (e.g., a stationary time series), this model will never do - since, on the other hand, Hume and many of his successors are not satisfied with any concrete framework. Given any such model, they say, in principle, things could be completely different tomorrow, beyond the scope of the model considered. So, no model will be appropriate - ever.

Table 1 Bounded vs. unbounded situations

1.	concrete model	specific prognosis
2.	no framework	no sustained prognosis

Given this, i.e., without any boundary conditions, restricting the situation somehow, we are outside *any* framework. But without grounds, nothing at all can be claimed, and principal doubt indeed is justified. However, in a sense, this is not fair or rather trivial: *Within* a reasonable framework, i.e., given some adequate assumptions, sound

[5] Also note the elegant symmetry: A set of 'top down' boundary conditions is equivalent to the set of all objects that adhere to all these conditions. Therefore substracting one of the boundary conditions is equivalent to extending the given set of objects to the superset of all those objects adhering to the remaining boundary conditions.

[6] An analogy can be found in the universe of sets which is also ordered hierarchically in a natural way, i.e., with the help of the subset operation ($A \subseteq B$). Starting with an arbitrary set A, the more general union set $A \cup B$ is always defined (i.e., for any set B), and so is the inductive gap $B \backslash A$. However, the union of all sets U no longer is a set. Transcending the framework of sets, it also turns the gap $U \backslash A$ into an abyss.

conclusions are the rule and not the exception. Outside of any such model, however, reasonable conclusions are impossible. You cannot have it both ways, i.e., request a sustained prognosis (first line in table 1), but not accept any framework (second line in table 1). Arguing 'off limits' (more precisely, beyond any limit restricting the situation somehow) can only lead to a principled and completely negative answer.

It should be added that there is also a straightforward logical argument against a general inductive principle: A general law is strong, since it - deductively - entails specific consequences. Alas, since induction is the opposite of deduction, some general inductive principle (being the limit of particular inductive rules) would have to be *weaker* than any specific inductive step. Thus, even if it existed, such a principle would be exceedingly weak and would therefore hardly support anything.

5 Successful inductions

5.1 A paradigmatic large gap: Connecting finite and infinite sequences

Given a finite population of size n, forming the upper layer, and data on k individuals (those that have been observed), there is an inductive gap: The $n-k$ persons that have not been observed. Closing such a gap is trivial: just extend your data base, i.e., extend the observations to the persons not yet investigated. Now, since we are dealing with the problem in a formal way, statistics has no qualms dealing with infinite populations. The upshot of sampling theory is that, even in this case, a rather small, but carefully (i.e., randomly) chosen subset suffices to get 'close' to properties (i.e., parameters) of the population. It is the standard assumptions built into the statistical paradigm - though seldom mentioned explicitly - that guarantee the convergence of the sample (and its properties) toward the larger population.

Given this, one is back to studying the link between a finite sequence or sample $\mathbf{x}_n = (x_1, \ldots, x_n)$ on the one hand, and all possible infinite sequences $\mathbf{x} = x_1, \ldots, x_n, x_{n+1}, \ldots$, starting with $\mathbf{x}_n$ on the other. Without loss of generality, we may focus on finite and infinite *binary* strings, i.e., $x_i = 0$ or 1 for all i. Thus we have two well defined layers, and we are in the mathematical world. However, there is a fairly large gap to be filled.

Deterministic approach

Kelly (1996) uses the theory of computability (recursive functions) to bridge this enormous gap, already encountered by enumerative induction. His approach is mainly topological, and his basic concept is logical reliability. Since "logical reliability demands convergence to the truth on *each* data stream" (ibid., p. 317, my emphasis), his overall conclusion is rather pessimistic: "... classical scepticism and the modern theory of computability are reflections of the same sort of limitations and give rise to demonic arguments and hierarchies of underdetermination" (ibid., p. 160). In the worst case, i.e., without further assumptions (restrictions), the situation is hopeless (see, in particular, his remarks on Reichenbach, ibid., pp. 57-59, 242f).

Not surprisingly, he needs a strong regularity assumption, called 'completeness', to get substantial results of a positive nature (ibid., pp. 127, 243): "The characterization theorems...may be thought of as proofs that the various notions of convergence are complete for their respective Borel complexity classes...As usual, the proof may be viewed as a completeness theorem for an inductive architecture suited to gradual identification."

Probabilistic approach

Much earlier, de Finetti (1937) had realized that any finite sample contained too little information to pass to some limit without hesitation. In particular he and others challenged the third major axiom of probability theory: If $A_i \cap A_j = \emptyset$ for all $i \neq j$, nobody doubts finite additivity, i.e., $p(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} p(A_i)$. However, mathematicians needed, and indeed just assumed, *countable* additivity: $p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$.

Kelly (1996) finally gives a reason why the latter assumption has worked so well. Upon giving up *logical* reliability in favour of *probabilistic* reliability which "... requires only convergence to the truth over some set of data streams that carries sufficiently high probability" (ibid., p. 317), he realizes that induction becomes much easier to handle. In this (weaker) setting, countable additivity plays the role of a crucial regularity (continuity) condition, guaranteeing that most of the probability mass is concentrated on a *finite* set. In other words, because of this assumption, one may ignore the end piece $x_{m+1}, x_{m+2}, \ldots$ of any sequence in a probabilistic sense (ibid., p. 324). A 'nice' consequence is that Bayesian updating, being rather dubious in the sense of logical reliability (ibid., pp. 313-316), works quite well in a probabilistic sense.

By now it should come as no surprise that "the existence of a relative frequency limit is a strong assumption" (Li and Vitányi (2008), p. 52). Therefore it is amazing that classical random experiments (e.g., successive throws of a coin) straightforwardly lead to laws of large numbers. That is, if $\mathbf{x} = x_1, x_2, \ldots$ satisfies certain *mild* conditions, the relative frequency r_n of the ones in the sample $\mathbf{x}_n$ *converges* (rapidly) towards r, the proportion of ones in the population. Our overall setup explains why:

First, there is a well-defined, constant population, i.e., an upper tier (e.g., an urn with a certain proportion r of red balls; some well-defined population parameter, in general). Second, the distance between sample and population is bounded (see section 3.3). Third, random sampling connects the tiers in a consistent way. In particular, the probability that a ball drawn at random has the colour in question is r (which, if the number of balls in the urn is finite, is Laplace's classic definition of probability). Because of these assumptions, transition from some random sample to the population becomes smooth:

The set S_∞ of all binary sequences $\mathbf{x}$, i.e., the infinite sample space, is (almost) equivalent to the population (the urn). That is, most of these samples contain $r\%$ red balls. (More precisely: The subset of those sequences containing about $r\%$ red balls is a set of measure one.) Finite samples $\mathbf{x}_n$ of size n are less symmetric (Li and Vitányi (2008), p. 168), in particular if n is small, but no inconsistency occurs upon moving to $n = 1$, since the probability of a red ball turning up is r.

Conversely, let S_n be the space of all samples of size n. Then, since each finite sequence of length n can be interpreted as the beginning of some longer sequence, we have $S_n \subset S_{n+1} \subset \ldots \subset S_\infty$. Owing to the symmetry conditions just explained and the well-defined upper tier, increasing n as well as finally passing to the limit does not cause any pathology. Rather, one obtains a law of large numbers (convergence in probability towards a population parameter) or other limit theorems (convergence in distribution), in particular if higher moments of the random variables involved exist.

Kolmogorov's approach

It may be noted that contemporary information theory builds on Kolmogorov complexity $K(\cdot)$ rather than probability, not least since with this ingenious concept, the link between the finite and the infinite becomes particularly elegant: First, some sequence $\mathbf{x} = x_1, x_2, \ldots$ with initial segment $\mathbf{x}_n = (x_1, \ldots, x_n)$ is called algorithmically random if its complexity grows fast enough, i.e., if the series $\sum_n 2^n/2^{K(\mathbf{x}_n)}$ is bounded (Li and Vitányi (2008), p. 230). Second, $\mathbf{x}$ is algorithmically random *if and only if* "the complexity of each initial segment is at least its length." (ibid., p. 221). In plain English the theorem just stated says that one is able to move from random (very complex) finite vectors to random infinite series - and back - seamlessly: Both possess 'almost' maximum Kolmogorov complexity.[7]

Finally, it should be mentioned that the theory is able to explain the remarkable phenomenon of a "practical limit" or "apparent convergence" *without* reference to a (real) limit: Most finite binary strings have high Kolmogorov complexity, i.e., they are virtually incompressible. According to Fine's theorem (cf. Li and Vitányi (2008), pp. 141), the fluctuations of the relative frequencies of these sequences are small. Both facts combined explain "why in a typical sequence [...] the relative frequencies appear to converge or stabilize. Apparent convergence occurs because of, and not in spite of, the high irregularity (randomness or complexity) of a data sequence" (ibid., p. 142).

In a nutshell: A "practical limit" is the rule, but real convergence is the exception, since the latter property is a strong assumption (i.e. there are only relatively few sequences that have this property). Thus Reichenbach's 'pragmatic justification' of induction is a red herring, apparent convergence being thoroughly misleading.

5.2 Coping with circularity

All verbal arguments of the form "Induction has worked in the past. Is that good reason to trust induction in the future?" or "What reason do we have to believe that future instances will resemble past observations?" have more than an air of circularity to them. If I say that I believe in the sun rising tomorrow since it has always risen in the past, I seem to be begging the question. More generally: Inductive arguments are plagued by the reproach that they are, essentially, circular.

Given a sequential interpretation, suppose we use all the information $I(n)$ that has occurred until day n in order to proceed to day $n+1$. It seems to be viciously circular to use all the information $I(n-1)$ that has occurred until day $n-1$ in order to proceed to day n, etc. However, partial recursive functions (loops), very popular in computer programming, demonstrate that this need not be so:

$n!$, called "n factorial", is defined as the product of the first n natural numbers, that is $n! = 1 \cdot 2 \cdots n$, for instance $4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$. Now, consider the program

FACTORIAL[n]:

- IF $n = 1$ THEN $n! = 1$

[7] The complete theory is developed in Li and Vitányi (2008), chapters 2.4, 2.5, 3.5, and 3.6. Their chapter 1.9 gives a historical account, explicitly referring to von Mises, the 'father' of Frequentism. Since any regularity is a kind or redundancy that can be used to compress a given string of data, the basic idea is to identify incompressibility with randomness: (almost) incompressible $\approx$ highly complex $\approx$ algorithmically random.

- ELSE $n! = n\cdot$ FACTORIAL$[n-1]$

At first sight, this looks viciously circular, since 'factorial is explained by factorial.' That is, the factorial function appears in the definition of the factorial function, and it is used to calculate itself. One should think that, for a definition to be proper, at the very least, some object (such as a specific function) *must not* be defined with the explicit help of this very object. Yet, as a matter of fact, the second line of the program implicitly defines a loop that is evoked just a finite number of times. For instance,

$$4! = 4 \cdot 3! = 4 \cdot 3 \cdot 2! = 4 \cdot 3 \cdot 2 \cdot 1! = 4 \cdot 3 \cdot 2 \cdot 1 = 24.$$

The point is that, on closer inspection, the factorial function depends on n. Every time the program FACTORIAL$[\cdot]$ is evoked, a different argument is inserted, and since the arguments are natural numbers descending monotonically, the whole procedure terminates after a finite number of steps. Shifting the problem from the calculation of $n!$ to the calculation of $(n-1)!$ not only defers the problem, but also makes it easier. Because of boundedness, this strategy leads to a well-defined algorithm, calculating the desired result.

Reversing the direction of thought, one encounters a self-similar expanding structure: Based on $1! = 1$, one proceeds to the next level with the help of $n! = n \cdot (n-1)!$ Thus, in a nutshell, a partial recursive function is an illustrative example of a benign kind of self-reference (circularity). The loop it defines is *not* a perfect circle, but a spiral with a well-defined starting point, and rotations that build on each other. In this picture, "times n" means to add another similar turn or to enlarge a given shape:

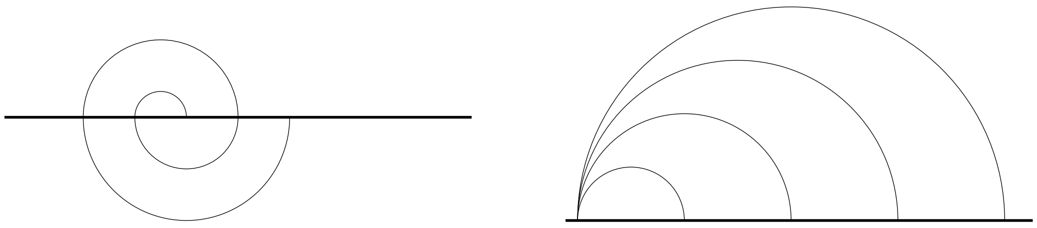


Fig. 3 Partially recursive structures

One may also interpret such a setting as an inductive funnel. In that picture, "times n" means to add another similar layer:

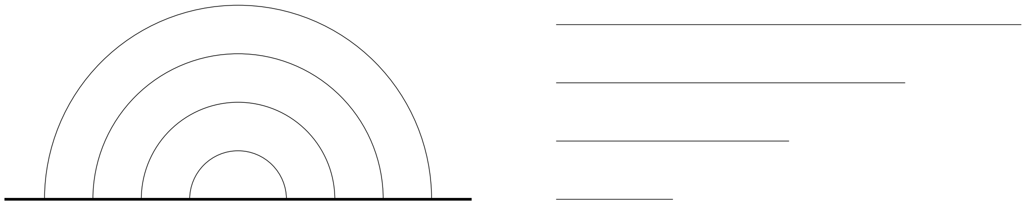


Fig. 4 Hierarchal structures

Either interpretation shows that the seemingly vicious loop entails a well-defined layered design. Quite similarly, set theory is grounded in the empty set and forms a cumulative hierarchy, in particular, since the axiom of regularity rules out circular dependencies.

The latter perspective demonstrates that using the information $I(n)$ in order to get to $I(n+1)$ need not be viciously circular. If $I(n) < I(n+1)$ for all natural numbers n, there is no 'vicious' circularity, but rather a many-layered hierarchy. Thus, it is not a logical fallacy to evoke "the sun has risen n times" in order to justify "the sun will rise $n+1$ times." Only if the assumptions were at least as strong as the conclusions, i.e., in the case of a tautology or a deductive argument, would one be begging the question.

However, we are dealing with induction, i.e., the fact that the sun has risen until today does *not* imply that it will rise tomorrow. In other words, there are inevitable inductive gaps, a sequence of assumptions that become weaker and weaker: $I(n)$ being based on $I(n-1)$, being based on $I(n-2)$, etc. This either leads to a finite chain (e.g., $I(1)$ representing the first sunrise), or to a sound convergence argument (discussed in subsection 4.2), since information is non-negative.

Starting with day 1, instead, evidence accrues. That is, $I(1) \leq I(2) \leq \ldots \leq I(n)$, and $I(n)$ may be considerably larger than $I(1)$. Of course, if the gaps $I(k+1) - I(k)$ are large, this way to proceed might not necessarily be convincing, but that is a different question.

Quantitative considerations

An appropriate formal model for the latter example is the natural numbers (corresponding to days $1, 2, 3$, etc.), having the property S_i that the sun rises on this day. That is, $S_i = 1$ if the sun rises on day i, and $S_i = 0$ otherwise.

Given the past up to day n with $S_i = 1$ for all $i \leq n$, the inductive step consists in moving to day $n+1$. In a sense, i.e., without any further assumptions, we cannot say anything about S_{n+1}, and of course $S_{n+1} = 0$ is possible. However, if we use the sequential structure and acknowledge that information accrues it seems to be a reasonable idea to define the inductive gap, when moving from day n to day $n+1$, as the fraction

$$\frac{n+1}{n} = 1 + \frac{1}{n}$$

In other words, if each day contributes the same amount of information (1 bit), $1/n$ is the incremental gain of information. Quite obviously, the relative gain of information goes to 0 if $n \to \infty$.

The above may look as though the assumption of "uniformity of nature" (associated with circularity) had to be evoked in order to get from day n to day $n+1$. However, this is not so. First, since the natural numbers are a purely formal model, some natural phenomenon like the sun rising from day to day is merely used as an illustration. Second, although the model can be interpreted in a sequential manner one need not do so. One could also say: Given a sample of size n, what is the relative gain of adding another observation? The answer is that further observations provide less and less information relative to what is already known. It is the number of objects already observed that is crucial, making the proportion larger or smaller: Adding 1 observation to 10 makes quite an impact, whereas adding 1 observation to a billion observations does not seem to be much.

Laplace (1812), who gave a probabilistic model of the sunrises, argued along the same lines and came to a similar conclusion. Since every sunrise contributes some information, the conditional probability that the sun will rise on day $n+1$, given that it has already risen n times should be an increasing function in n. Moreover, since the relative gain of information becomes smaller, the slope of this function would decrease monotonically. Laplace's calculations yielded $p(S_{n+1}|S_1 = \ldots = S_n = 1) = (n+1)/(n+2)$ which, as desired, is an increasing and concave function in n. This model only uses the information given by the data (i.e., the fact that the sun has risen n times in succession). He noted that the probability becomes (much) larger, if further information about our solar system is taken into account.

Asymptotic information theory

One might object that the physical principle of "uniformity of nature" is replaced by the (perhaps equally controversial) formal principle of "indifference", i.e., that each observation is equally important or carries the same "weight" $w \geq 0$. But that is not true either. For even if the observations carry different weights $w_i \geq 0$, the conclusion that further data points become less and less important remains true *in expectation*: Given a total of n observations, their sum of weights w.l.o.g. being 1, the expected weight of each of these observations has to be $1/n$. Thus we expect to collect the information $I(k) = k/n$ with the first k observations, and the expected relative gain of information that goes with observation $k+1$ still is $\Delta(k) = I(k+1)/I(k) = \frac{k+1}{n}/\frac{k}{n} = \frac{k+1}{k} = 1 + 1/k$.

To make $\Delta(n)$ large, one would have to arrange the weights in ascending order. However, 'save the best till last' is just the opposite to some typical scenario that can be expected for reasons of combinatorics. Uniformity in the sense of an (approximate) uniform distribution is a mere *consequence* of these considerations due to the asymptotic equipartition property (see, e.g., Cover and Thomas (2006)), chap. 3. Rissanen (2007), p. 25, explains:

> As n grows, the probability of the set of typical sequences goes to one at the near exponential rate ... Moreover ... all typical sequences have just about equal probability.

It may be added that "non-uniformity" in the sense that the next observation could be completely different from all the observations already known cannot occur in the above model, since every observation only has a finite weight w_i which is a consequence of the fact that the sum of all information is modelled as finite (w.l.o.g. equal to 1).

In general, i.e., *without* such a bound, the information $I(A)$ of an event A is a monotonically decreasing and continuous function of its probability. Thus if the probability is large, the surprise (amount of information) is little, and vice versa. In the extreme, $p(A) = 1 \Leftrightarrow I(A) = 0$ (some sure event is no surprise at all), but also $p(A) = 0 \Leftrightarrow I(A) = \infty$. That is, if something deemed impossible happens, $I = \infty$ indicates that the framework employed is wrong.

5.3 Invertibility

Throughout this contribution, and perfectly in line with the received framework, inductive problems involve (at least) two levels of abstraction being connected in an *asymmetric* way. If these levels are coupled together with the help of a certain operation, the classical issue of inferring a general pattern from the observation of particular instances translates into models consisting of two layers and an asymmetric relation between them. Consistently, the most primitive of such models is defined by two sets A, C, connected by the subset relation $\subseteq$. The corresponding operation that simplifies matters (wastes information) is a non-injective mapping $f : A \to C$. In other words, there are elements $a \in A$ having the same image $c = f(a)$ in C, and the size of the set of all members of A that are mapped to an 'non-exclusive' $c \in C$ is a natural measure of the mapping's invertibility. In the extreme, all $a \in A$ are mapped to a single c, so that, given c it is impossible to say anything about this observation's origin.

Neglecting the layers and focussing on the operation, it is always possible to proceed from the more abstract level (containing more information) to the more concrete situation (containing less information). This step may be straightforward or even trivial. Typically, the corresponding operation simplifies matters and there are thus general rules governing this step (e.g., adding random bits to a certain string, conducting a random experiment, executing an algorithm, differentiating a function, making a statement less precise, etc.). Taken with a pinch of salt, this direction may thus be called *deductive*.

Yet the inverse *inductive* operation is *not* always possible or well-defined. It only exists sometimes, given certain additional conditions, in specific situations. Even if it exists, it may be impossible to find it in a rigorous or mechanical way. For example, there is no general algorithm to compress an arbitrary string to its minimum length, objects can exist but may not be constructible, and it can be very difficult to find a causal relationship behind a cloud of correlations. Consistently, Bunge (2019) emphasizes that "inverse problems" are much more difficult than "forward problems."

In general, there is a *continuum of reversibility*: One extreme is perfect reversibility, i.e., given some operation, its inverse is also always possible. For example, + and - are perfectly symmetric. If you can add two numbers, you may also subtract them. That is not quite the case with multiplication, however. On the one hand, any two numbers can be multiplied, but on the other hand, all numbers except one (i.e., the number 0) can be used as a divisor. So, there is almost perfect symmetry with 0 being the exception. Typically, an operation can be *partially* inverted. That is, f^{-1} can be explained given some special *conditions*. These conditions can be non-existent (any differentiable function can be integrated), mild (division can almost always be applied), or quite restrictive (in general, a^b is only defined for positive a).

Thus, step by step, we arrive at the other extreme: perfect non-reversibility, i.e., an operation cannot be inverted at all. For example, given a sequence $\mathbf{x} = x_1, x_2, x_3, \ldots$, it is trivial to proceed to $\mathbf{x}_n = (x_1, \ldots, x_n)$, since one simply has to skip $x_{n+1}, x_{n+2}, \ldots$ However, without further assumptions, it is impossible to infer anything about $\mathbf{x}_n$'s succession. Although the operation (link) between $\mathbf{x}$ and $\mathbf{x}_n$ is just a well-defined projection, it is also a so-called "trapdoor function." That is, having traveled through this door, it is impossible to get back, since the distance between the finite and the infinite situations (the floor and the ceiling so to speak) is unbounded.

Cases near the latter pole lend credibility to Hume's otherwise amazing claim that only deduction can be rationally justified, or that induction does not exist at all (Popper). Yet there is a large middle ground held by "partial invertibility." For example, Knight (1921), p. 313, says:

> The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past.

Probability theory and statistics

Upon trying to solve Hume's problem, many philosophers - most notably Reichenbach, Carnap, Popper, and the Bayesian school (Howson und Urbach 2006) - have looked to probability and statistics. A major reason could be that invertibility is quite straightforward in that area:

Given some set Ω, the first axiom of probability states that $p(\Omega) = 1$, i.e., that the total probability mass is bounded. Therefore, if you know the probability $p(A)$ of some event A, you may straightforwardly compute the probability of the opposite event $\bar{A}$, since $p(\bar{A}) = 1 - p(A)$.

In statistics, a standard way to encode various hypothetical laws is by means of a parametric family of probability distributions $p_\theta(x)$, leading to the *lines* of the following exemplary matrix:

Table 2 Probability distributions and likelihood

	x_1	x_2	x_3	x_4	$\sum$
θ_1	0.2	0.3	0.1	0.4	1
θ_2	0.5	0.3	0.2	0	1
θ_3	0.25	0.25	0.3	0.2	1

Given the observation $X = x_4$, say, a guess $\hat{\theta}$ of the true (but unknown) θ is straightforward: Just switch to the fourth *column* and choose the "maximum likelihood" there, that is: $L_{x_4}(\theta) = p_\theta(x_4) = (0.4, 0, 0.2)$, $max\ L_{x_4}(\theta) = 0.4$, and thus $\hat{\theta} = \theta_1$. Moreover, due to the observation of x_4, one may exclude the hypothetical value θ_2.

The Bayesian framework extends this reasoning upon introducing prior probabilities $q(\theta) = (q(\theta_1), q(\theta_2), q(\theta_3))$ and bases its inferences on the posterior distribution $q(\theta|x) = (q(\theta_1|x), q(\theta_2|x), q(\theta_3|x))$. For example, if $q(\theta_i) = 1/3$ for $i = 1, 2, 3$, Bayes' formula gives the inverse probabilities $q(\theta_2|x_4) = 0$, and

$$q(\theta_1|x_4) = \frac{0.4/3}{(0.4+0.2)/3} = \frac{2}{3} = 1 - \frac{1}{3} = 1 - \frac{0.2/3}{(0.4+0.2)/3} = 1 - q(\theta_3|x_4)$$

Thus, in a sense (and just as Fisher had claimed), the inductive step boils down to an elementary calculation.

It may be added that mathematics has found yet another way of dealing with the basic asymmetry: Owing to the non-injectivity of f, the inverse mapping straightforwardly leads to a larger (more general) class of objects. When the Greeks tried to invert multiplication $(a \cdot b)$, they had to invent broken numbers a/b, thus leaving behind the familiar realm of whole numbers. Inverting $a^2 = b$ led to roots, and thus the irrationals. If, in the last equation, b is a negative number, another extension becomes inevitable (the imaginary numbers, such as $i = \sqrt{-1}$).

6 Goodman's challenge

Stalker (1992) gives a concise description of Goodman's idea

> Suppose that all emeralds examined before a certain time t are green. At time t, then, all our relevant observations confirm the hypothesis that all

> emeralds are green. But consider the predicate 'grue' which applies to all things examined before t just in case they are green and to other things just in case they are blue. Obviously at time t, for each statement of evidence asserting that a given emerald is green, we have a parallel evidence-statement asserting that that emerald is grue. And each evidence-statement that a given emerald is grue will confirm the general hypothesis that all emeralds are grue [...] Two mutually conflicting hypotheses are supported by the same evidence.

In view of the funnel-structure discussed throughout this contribution, this bifurcation is not surprising. However, there is more to it:

> And by choosing an appropriate predicate instead of 'grue' we can clearly obtain equal confirmation for any prediction whatever about other emeralds, or indeed for any prediction whatever about any other kind of thing.

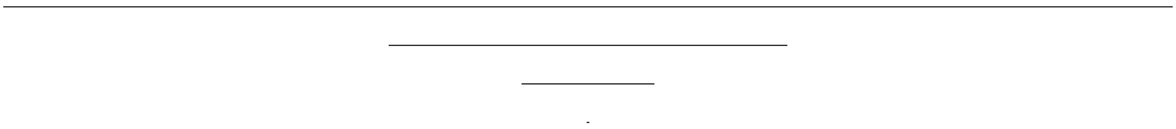
In other words, instead of criticizing induction like so many of Hume's successors, Goodman's innovative idea is to *trust* induction, and to investigate what happens next. Unfortunately, induction's weakness thus shows up almost immediately: It is not at all obvious in which way to generalize a certain statement - which feature is 'projectable' which is not (or to what extent)? For example, this line of argument may "lead to the absurd conclusion that no experimenter is ever entitled to draw universal conclusions about the world outside his laboratory from what goes on inside it." (ibid.)

In a nutshell, if we do not trust induction, we are paralysed, not getting anywhere. However, if we trust induction, this method could take us anywhere, which is almost equally disastrous. In our basic model, these cases correspond to $d(A, C) = 0$, and $d(A, C) = \infty$, respectively.

6.1 Boundedness is crucial

Figures 1 and 2 give a clue as to what happens: A sound induction is justified if the situation is bounded. So far, we have just looked at the y-axis, i.e., we went from a more specific to a more general situation. Since both levels are well-defined, a finite funnel is the appropriate geometric illustration. Goodman's example points out that one must also avoid an unbounded set A. In other words, there also has to be boundedness with respect to the x-axis. Geometrically, one must avoid the following situation, where the uppermost line is *unbounded*:

Fig. 5 Set Divergence



Although the inductive gap with respect to the y-axis is finite, the sets involved may diverge. In Goodman's example there isn't just a well-defined bifurcation or some restricted funnel. Instead, the crux of the above examples is that the specific situation is generalized in a wild, rather arbitrary fashion. (Just note the generous use of "any.") Considering GRUE: The concrete level C consists of the constant colour green, i.e., a single point. The abstract level A is defined by the time t a change of colour occurs. This set has the cardinality of the continuum and is clearly unbounded. It would suffice to choose the set of all days in the future (1 = tomorrow, 2 = the day after tomorrow, etc.) indicating when the change of colour happens, and to define $t = \infty$ if there is no change of colour. Clearly, the set $I\!N \cup \{\infty\}$ is also unbounded. In both models we would not know how to generalize or why to choose the point ∞.

Here is a practically relevant variation: Suppose we have a population and a large (and thus typical) sample. This may be interpreted in the sense that, with respect to some *single* property, the estimate $\hat{\theta}$ is close to the true value θ in the population. However, there is an infinite number of (potentially relevant) properties, and with high probability, the population and the sample will differ considerably in at least one of them. Similarly, given a large enough number of nuisance factors, at least one of them will thwart the desired inductive step from sample to population, the sample not being representative of the population in this respect.

The crucial point, again, is boundedness. Boundedness must be guaranteed with respect to the sets involved (the $x-$axis), *and* all dimensions or properties that are to be generalized.[8] In Goodman's example this could be the time t of a change of colour, the set of all colours c taken into account, or the number m of colour changes. Thus the various variants of Goodman's example point out that inductive steps are, in general, *multidimensional.* Several properties or conditions may be involved and quite a large number of them may be generalized *simultaneously.* Geometrically speaking, the one-dimensional inductive funnel becomes a multi-dimensional (truncated) pyramid.

In order to make a sound inductive inference, one firstly has to refrain from arbitrary, i.e., unbounded, generalizations. Replacing "colour" by "any predicate", and "emeralds" by "any other thing" inflates the situation beyond any limit. Introducing a huge number of nuisance variables, an undefined number of potentially relevant properties, or infinite sets of objects may also readily destroy even the most straightforward inductive step. Stove (1986), p. 65, is perfectly correct when he remarks that this is a core weakness of Williams' argument. In the following quote (cf. Williams (1947), p. 100) summarizing his ideas, it's the second *any* that does the harm:

> Any sizeable sample very probably matches its population in any specifiable respect.

Given a fixed sample, and an infinite or very large number of 'respects', the sample will almost certainly not match the population in at least one of these respects. However, given a certain (fixed) number of respects, a sample will match the population in all of these respects if n is large enough. By the same token, Rissanen concludes that a finite number of observations allows one to distinguish between a *finite* but not an arbitrary number of hypotheses.

6.2 Levels of abstraction

Inductive problems appear in various guises. First, we considered two tiers and studied the distance between them. Second, we focused on the mapping from A to C and its

[8] If an object has a certain property, it complies with a certain (rather strict) constraint. So, in principle, generalizing a particular property is no different from weakening a certain condition.

inverse. While the symmetric concept of *distance* highlights the similarity of A and C; the inverse function, logical implication, and the subset-relation are all asymmetric, and thus point at the difference.[9]

Although such a clear separation is more transparent than some notion of "partial invertibility" that easily confounds both perspectives, Goodman's challenge hints at another, a third class of problems: Given a concrete sample C - what could be a reasonable generalization A (plus a natural mapping connecting these tiers)?

In other words: Starting with some piece of information, very often the most serious problem consists in finding a suitable level of abstraction. Having investigated some platypuses in zoos, it seems to be a justified conclusion that they all lay eggs (since they all belong to the same biological species), but not that they all live in zoos (since there could be other habitats populated by platypuses). In Goodman's terms, *projectibility* depends on the property under investigation. Depending on the specific situation, it may be non-existent or (almost) endless. The stance taken so far is that as long as there is a bounded funnel-like structure in *every* direction of abstraction, the inductive steps taken are rational. The final result of a convincing inductive solution always consists of the sets (objects), dimensions (properties) and boundary conditions involved, plus (at least) two levels of abstraction in each dimension.

Given just C, however, the difficulty is twofold: First, one always has to construct A, i.e., A is not given. Second, upon constructing A, there is the problem of determining a reasonable distance $d(A,C)$. On the one hand, a critical attitude (distrust) easily leads to $d(A,C) = 0$. On the other hand, an optimistic point of view (trust in induction) straightforwardly results in $d(A,C) = \infty$. Moreover, in the worst case scenario, just a single finite $d(A,C)$ will do - we should claim neither more nor less. This is as subtle a problem as sending a satellite into orbit: If the impetus is too weak, it will come down again; if the impetus is too strong, it will disappear.

Thinking of "projectibility" and its opaqueness, there may be specific answers in certain scenarios. But be this as it may, is there a general answer without reference to a semantic context? In section 3.2, the discussion of Rissanen's example already hinted at the general solution given by the information sciences. Informally,

$$\text{Prior information} + \text{Information in some set of data} = \text{Posterior information}$$

Formally, the information of some event is just the negative logarithm of its probability. Therefore adding information is tantamount to multiplying probabilities. If the events are a hypothesis H and data $\mathbf{x}$ we get

$$\begin{aligned} I(H,\mathbf{x}) = I(H) + I(\mathbf{x}|H) \quad &\Leftrightarrow \quad -\log p(H,\mathbf{x}) = -\log p(H) - \log p(\mathbf{x}|H) \\ &\Leftrightarrow \quad p(H,\mathbf{x}) = p(H) \cdot p(\mathbf{x}|H) \end{aligned} \tag{1}$$

In other words: The first tier C consists of the prior hypothesis H, the second tier A is the more detailed information available after having seen the data $\mathbf{x}$, consisting of H and $\mathbf{x}$. The difference is just the information in the data which is not already contained in H. (Things that are known do not need to be learned twice.) The step from prior to posterior is inductive in the sense that a vague picture becomes more

[9] Quite similarly, in information theory, mutual information $I(A,C)$ is a fundamental symmetric concept, and KL-divergence, being at least as important, is asymmetric (Cover and Thomas 2006).

detailed. Notice, that (1) does not generalize from the data to the hypothesis, but rather from the hypothesis to the hypothesis plus the data.

Qualitatively speaking, it is not obvious how much “distance” is bridged by the data. If the distance $I(\mathbf{x}|H)$ were too small, we would not use all the information available. Thus we would lose if we played against somebody using all the information at hand. If $I(\mathbf{x}|H)$ were too large, we would go too far, i.e., a part of the conclusion could not be substantiated with the help of the data, amounting to a real inductive (non-deductive) step. Thus, somebody using just the amount of information truly available - but no more - would also beat us if we gambled.

In other words, there is a *unique* logically sound solution: Prior information $I(H)$ and the information in the data conditional on the prior $I(\mathbf{x}|H)$ must add up to the total information $I(H, \mathbf{x})$ which is equation (1). Traditionally, the gambling schemes just introduced are known as Dutch books. It is only possible to avoid them if the decisions of a gambler are consistent, i.e., if he adheres to the axioms of probability theory (Ramsey 1926, de Finetti 1937, Greenland 1998) which is also reflected in (1). Finally, symmetrizing equation (1) immediately yields Bayes’ formula

$$p(B|A) \cdot p(A) = p(A, B) = p(A|B) \cdot p(B) \Leftrightarrow p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

with general events A, B. Therefore it seems to be no coincidence that Bayesians are among the most ardent defenders of induction.

Note that Rissanen’s notion of ‘*optimal* distinguishability’ implements the same idea: Given some set of data, i.e., a certain amount of information, there is a corresponding maximum number of rationally distinguishable hypotheses. In a simplistic manner, one could say that one needs more/less data in order to differentiate between more/fewer hypotheses.

7 Summary

In sum, at least formally, Hume’s problem can be dealt with in a constructive way. It turns out that the key concepts are boundedness and information. If the mutual information or distance between A and C is larger than zero, it is possible to say something about C given A - and vice versa! To this end, there has to be a link between A and C. In theoretical settings the logical link necessary is provided by some assumption. Hume (1748/2008) was right in noticing that a strong condition like ‘resemblance’ may easily be violated. However, a weak kind of concatenation may also suffice for an educated guess.

In principle almost any kind of coupling will do, *as long as the link transports some information.* Here is a straightforward counterexample: If information flows from C to A but not from A to C (i.e., given a trapdoor function or an invisible observer in the sky), A is able to learn something about C, but not vice versa. Arguably, the most straightforward way to model a linkage between two sets is the subset relation, i.e., $C \subseteq A$, or, almost equivalently, to consider a sample C *from* some population A. In the easiest case, it is just the cardinality of the sets that is relevant. More precisely, if $|M|$ denotes the cardinality of some set M, and $C \subseteq A$, the larger the ratio $|C|/|A|$, the more one knows about A, given the sample C. In particular, if $A = C \Rightarrow |C|/|A| = 1$, there is no inductive problem; if $|C| = 0$, i.e., if there is no sample from A, nothing can be said about A. Enumerative induction is not convincing since in this case $|C|$ is a finite number, but $|A| = \infty$. However, as long as the ratio $|C|/|A|$ remains finite, it seems fair to say that there is some evidence in C about A, and the closer $|C|/|A|$ is to one, the stronger this evidence is.

A more sophisticated model should distinguish between cardinality and information. The most straightforward formalization could be $C \subseteq A$ and $I(C) < I(A)$. Given this situation, there are three ways to proceed:

(i) Deductively, i.e., from A to C.

(ii) Inductively, i.e., from C to A, leaping over a gap of size $I(A) - I(C)$.

(iii) Additively, i.e., $I(C) + d = I(A)$, with $d > 0$ bridging the gap, see equation (1).

More generally speaking, a model should formalize $I_A(C)$, the information in C about some property or parameter of A. In statistics, C is represented by a sample and A is some population. Bad samples and difficult estimation problems have in common that $I_A(C)$ is low, although the size of the sample might be considerable. In the worst case, $I_A(C)$ is minute, despite the fact that $|C|$ and $|A|$ do not differ much.[10] This is where random sampling comes in: For combinatorial reasons, typically $I_A(C)$ is large, although the size of the sample may still be rather small.

All in all, statistics provides the most comprehensive response to Hume's problem: Given its basic sample-population model, standard assumptions guarantee convergence (LLN etc.) Given the link between probability and information, one may calculate how to proceed from C to A (information adds up, probabilities multiply, and Bayesian updating is the standard way to include new information). Finally, the axioms of probability provide a sound foundation for probability calculations, exclude Dutch books (inconsistency) and underpin rationality. On top of this, countable additivity essentially reduces infinite sequences to finite samples. Thus these considerations combined pin down the crucial ingredients of the problem and provide constructive answers.

Alas, it is quite typical for verbal discussions that they easily miss their target. The whole philosophical discussion seems to revolve around the "rhetoric of independence" (cf. Stove (1986), Ch. VIII). That is, at the end of the day, all arguments in favour of induction tend to be unconvincing, since the observed and the unobserved are "loose and separate", "disjoint", "lack connection", etc. However, consider deterministic chaos: Although there exists a deterministically strong connection (i.e., if you know the present state x of the system, you also *know* its future state $y = f(x)$), the mapping f also scatters the points of any neighbourhood $U(x)$ of x so effectively that prediction becomes a futile endeavour. So, in a sense, the link between the present and the future is strong and weak simultaneously, and one may easily bark up the wrong tree.

Instead, the crucial point - again - is unboundedness: The information at present, no matter how detailed, decays exponentially, leaving no ground for a rational prediction.

[10] If every observation contributes (about) the same amount of information, all samples of size n are equivalent in the sense that they contain (about) the same amount of information. In other words, in such 'nice' situations, quite typical for classical statistics, there are neither good nor bad samples. All that matters is the size n of the sample (the larger n the better). This is not so in difficult estimation problems, when a few of the observations contain most of the information. Thus information does not accrue steadily. Rather, there are a few sudden jumps. In the most extreme case, all depends on a single and thus crucial observation.

Equivalently, one can say that uncertainty (errors) grows very fast, soon exceeding any bound. In this view, Hume's stance amounts to the remarkable claim that the 'predictive horizon' is precisely zero, no exception. In other words: Scepticism in its most extreme form (i.e., rational prediction is strictly impossible) is tantamount to every bit of information always vanishing instantaneously.

However, Hume was right that there is no such thing as a free lunch (NFL). Without any assumptions, boundary conditions or restrictions - beyond any bounds - there are no grounds for induction. Indeed, it has been possible to prove various "NFL theorems" in the field of machine learning (Wolpert 2013), for instance "... just because Professor Smith produces search algorithms that outperform random search in the past, without making some assumption about the probability distribution over universes, we cannot conclude that they are likely to continue to do so in the future" (ibid., p. 7).

Induction only works within certain frameworks, e.g., those of statistics. Groarke (2009), p. 61, makes a similar point: "Yes, post-Cartesian scepticism undermines inductive reasoning, but it also undermines deduction and everything else. This kind of absolute doubt, consistently practiced, makes genuine knowledge impossible." It may be conceded, however, that crucial assumptions are often kept implicit, thus we had to dig deep into technical details to expose them. Without these boundary conditions, the information sciences - at least their asymptotic theories - would not work.

For more on these matters see Saint-Mont (2017). Apart from being more comprehensive, that work also discusses many historical attempts and the empirical side to induction. In particular: why are we "dealing with an 'inductively normal' world - a world in which induction is an actually efficient mode of inquiry" (Rescher (1980), pp. 160-161)? Why are inductive leaps, typically, 'predictively reliable' or why do they possess, in Peirce's words, a 'truth-producing virtue' (cf. Braithwaite (1953), p. 264)?

Following the above train of thought, the roundabout answer seems to be that persisting patterns are the rule. That is, in the world we inhabit and in most of our models, information decays rather slowly (Costa (2018), pp. 357-359); and if information is at least conserved to some extent, predictions are feasible. Acknowledging that objects, forms and regularities are particular and particularly important expressions of information, it is *structure* - which may be understood as preserved information - rather than conformity that is crucial: One can have reasonable doubts in nature's uniformity, however, hardly anybody will challenge the fact that we live in a structured universe.

In sum, inductive steps can often be justified, *boundedness* and *information* being the crucial ideas, essentially leading to 'inductive funnels' (see the illustrations). In particular, convergence and (partially) recursive functions demonstrate that induction is far from worthless circular reasoning. Given any reasonable framework, inductive steps can be vindicated, i.e., Hume's remarkable claim that no inductive step is ever justified seems to be hardly more than an 'old wives' tale.' Reality chooses to be with cautious hierarchal non-circular inductive steps, rather than with principled doubt.

References

Baum, L.E.; Katz, M.; and R.R. Read (1962). Exponential convergence rates for the law of large numbers. Trans. Amer. Math. Soc. 102, 187-199.

Black, M. (1958). Self-supporting inductive arguments. J. of Phil., 55(17), 718-725.

Braithwaite, R.B. (1953/68). Scientific explanation. Cambridge: At the university press. Partly reprinted as chap. 7 in Swinburne (1974), 102-126.

Broad, C.D. (1952). The philosophy of Francis Bacon. In: Ethics and the history of philosophy. London: Routledge and Kegan Paul.

Bunge, M. (2019). Inverse problems. Foundations of Science, 24, 483-525.

Campbell, S. (2001). Fixing a Hole in the Ground of Induction. Australasian Journal of Philosophy, 79(4), 553–563.

Campbell, S.; and J. Franklin (2004). Randomness and the Justification of Induction. Synthese, 138(1), 79–99.

Costa, C. (2018). Philosophcial Semantics. Reintegrating theoretical philosophy. Newcastle upon Tyne: Cambridge Scholars Publishing.

Cover, T.M.; and Thomas, J.A. (2006). Elements of Information Theory. (2nd ed.) New York: Wiley.

Finetti, B. de (1937). La Prévision: ses Lois Logiques, ses Sources Sujectives. *Ann. Inst. H. Poincaré* **7**, 1-68.

Encylopedia Britannica (2018). Induction. www.britannica.com/topic/induction-reason

Érdi, P. (2008). Complexity explained. Berlin, Heidelberg: Springer.

Fazekas, I.; and O.I. Klesov (2000). A general approach to the strong laws of large numbers. Teor. Veroyatnost. i Primenen. 45(3), 568–583; Theory Probab. Appl., 45(3) (2002), 436–449.

Fisher, R.A. (1966). The Design of Experiments. (8th ed.). New York: Hafner Publishing Company.

Gauch, H.G. Jr. (2012). Scientific method in brief. Cambridge: Cambridge Univ. Press.

Greenland, S. (1998). Probability Logic and Probabilistic Induction. *Epidemiology* **9(3)**, 322-332.

Godfrey-Smith, P. (2003). Theory and Reality. Chicago and London: The University of Chicago Press.

Groarke, L. (2009). An Aristotelean account of induction. Montreal: McGill-Queen's University Press.

Hacking, I. (2001). An Introduction to Probabilty Theory and Inductive Logic. Cambridge: Cambridge University Press, Cambridge.

Howson, C.; and Urbach, P. (2006). Scientific Reasoning. The Bayesian Approach (3rd ed.). *Open Court, Chicago and La Salle, IL.*

Hume, D. (2008). 1st ed. 1748. An Enquiry Concerning Human Understanding. New York: Oxford University Press.

Indurkhya, Bb. (1990) Some Remarks on the Rationality of Induction. Synthese 85(1), 95–114.

Jaynes, E.T. (2003). Probability Theory. The Logic of Science. Cambridge: Cambridge University Press.

Kelly, K.T. (1996). The logic or reliable inquiry. New York, Oxford: Oxford Univ. Press.

Kemeny, J.G. (1953). The use of simplicity in induction. Phil. Review **62**, 391-408.

Knight, F. (1921). Risk, Uncertainty, and Profit. New York: Houghton Mifflin, New York.

Kullback, S. (1959). Information Theory and Statistics. *Wiley, New York.*

Laplace, P.-S. (1812). Théorie Analytique des Probabilités. Paris: Courcier.

Li, M.; and P. Vitányi (2008). An Introduction to Kolmogorov Complexity and its Applications. (3rd ed.) New York: Springer.

Papineau, D. (1992). Reliabilism, induction and scepticism. *Phil. Quart.* **42(166)**, 1-20.

Perrin, J. (1990). Atoms. Woodbridge, CT.: Ox Bow Press.

Popper, K.R.; and Miller, D.W. (1983). A proof of the impossibility of inductive probability. *Nature* **302**, 687-688.

Ramsey, F.P. (1926). Truth and Probability. In: Ramsey (1931), The Foundations of Mathematics and other Logical Essays, ch. VII (pp. 156-198), edited by Braithwaithe, R.B. *Kegan, Paul, Trench, Trubner & Co., London.*

Reichenbach, H. (1938). Experience and prediction. Chicago: University of Chicago Press.

Reichenbach, H. (1949). The theory of probability. Berkeley, CA: The University of California Press.

Reichenbach, H. (1956). The rise of scientific philosophy. Berkeley, CA: The University of California Press.

Rescher, N. (1980). Induction. An essay on the justification of inductive reasoning. Oxford: Basil Blackwell Publisher.

Rissanen, J. (2007). Information and Complexity in Statistical Modelling. New York: Springer.

Saint-Mont, U. (2017). Induction: The glory of science and philosophy. http://philsci-archive.pitt.edu/13057/1/USM_Induction2017.pdf

Solomonoff, R. (1964). A Formal Theory of Inductive Inference, Parts I and II. *Information and Control* **7**, 1-22, 224-254.

Stalker, D. (1992). Grue! The New Riddle of Induction. Chicago: Open Court.

Stove, D.C. (1986). The Rationality of Induction. Oxford: Clarendon Press.

Swinburne, R. (ed., 1974). The justification of induction. Oxford: Oxford Univ. Press.

Tukey, J.W. (1961). Statistical and Quantitative Methodology. In: Trends in Social Science. Ray, D.P. (ed.) New York: Philosophical Library, 84-136.

Walker, S. (2003). On sufficient conditions for Bayesian consistency. Biometrika **90(2)**, 482-488.

Walker, S. (2004). New approaches to Bayesian consistency. Ann. of Stat. **32(5)**, 2028-2043.

Whitehead, A.N. (1926). Science and the modern world. New York.

Will, F.L. (1953). Will the Future Be Like the Past? In: Flew, A. (ed.), Logic and Language: Second Series. Oxford: Blackwell, 32-50.

Williams, D. (1947). The Ground of Induction. Cambridge, MA: Harvard University Press.

Woit, P. (2006). Not Even Wrong. The Failure of String Theory and the Continuing Challenge to Unify the Laws of Physics. Jonathan Cape.

Wolpert, D.H. (2013). What the no free lunch theorems really mean. Ubiquity, Vol. 2013, December 2013, DOI: 10.1145/2555235.2555237