Induction: The glory of science and philosophy

Uwe Saint-Mont, Nordhausen University of Applied Sciences

May 22, 2017

> Any hopeful, humanitarian, knowledgeable, and right culture depends on induction not merely in its parts, to justify its particular scientific inquiries and political inventions. It depends on induction altogether and in principle. (Williams (1947), p. 16)

Abstract. The aim of this contribution is to provide a rather general answer to Hume's problem, the well-known problem of induction. To this end, it is very useful to apply his differentiation between "relations of ideas" and "matters of fact", and to reconsider earlier approaches.

In so doing, we consider the problem formally (chap. 3), as well as empirically (chap. 4). Next, received attempts to solve the problem are discussed (chap. 5). The basic structure of inductive problems is exposed in chap. 6. Our final conclusions are to the positive, i.e., Hume's problem can be dealt with - solved - in a constructive way (chap. 7). More specifically, bounded generalisations can be justified, and the key to the solution is the concept of information.

Keywords

Hume's Problem; Induction; Inference

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." In particular, since a crucial feature of science are 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 problem in more detail. Traditionally, "many dictionaries define inductive reasoning as reasoning that derives general principles from particular/individual instances" (e.g., see the English Wikipedia on "Inductive reasoning"). 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 - no doubt. 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. Any theorem must be deduced from the axioms via logical implications. In this sense any theorem is always more specific than the whole set of axioms, its content is more restricted than the complete realm defined by the axioms. For example, 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.

Perhaps surprisingly, one finds claims to the contrary in standard contemporary philosophical texts. For example, Gustason (1994), p. 16, claims that "there are valid arguments that proceed from the specific to the general... He says: "Detroit is a large city. [Thus] Everyone who lives near Detroit lives near a large city."

Obviously, Detroit plus its vicinity is more general than the city of Detroit. However, the crucial point is that we do not extend any property of the city of Detroit or its inhabitants to some larger area (which would be a non-trivial inductive step). Further classes of inconclusive counterexamples are discussed in Groarke (2009), pp. 35-38, and in the Internet Encyclopedia of Philosophy (IEP 2015).

2 Hume's problem

Inductive steps proceed from the more specific to the less specific, whereas deduction moves from the more general to the less general (tautologies included). Thus deduction preserves truth, whereas a real inductive step is never sure. Since the classical formulation of Hume's problem is plainer, we address the received problem of justifying generalizing conclusions. Let us start with his own words (Hume (1748/2008), Sect. IV., § 32):

> ...all inferences from experience suppose, as their foundation, that the future will resemble the past, and that similar powers will be conjoined with similar sensible qualities. If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion. It is impossible, therefore, that any arguments from experience can prove this resemblance of the past to the future; since all these arguments are founded on the supposition of that resemblance.

A concise modern exposition of the problem and its importance can be found in Gauch (2012), pp. 168:

> (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,[1] 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.

[1] "Deduction and induction are the only options" in Gauch's words

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 (see the very first quotation).

(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 attempt to justify induction. 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."

3 Formal treatment

Information theory possesses a very general and yet extremely simple framework within which the issue can be analysed and understood. Then we turn to mathematics, the science of abstraction.

3.1 First model

The basic unit of information is the Bit. This logical unit may assume two distinct values (typically named 0 and 1). Now, one needs to distinguish between two situations: 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.

Consider the following example:

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

Reading this table 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 nontrivial 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 wildcard ? 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.

[Illustration 1: Trapezoid with the shorter side at the bottom]

Split of 2. digit	00	01	10	11	
Split of 1. digit		$0,0$		$1,0$	
Starting sequence			$0,0$		

A refined model: distributions

Most authors who are concerned with 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, if you flip a coin, this may be modelled by a random variable assuming the values 1 (Heads) and 0 (Tails) with probabilities p and $1 - p$, respectively. (In statistical jargon, this is a Bernoulli distribution $B(p)$ with parameter p.) 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.

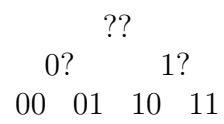
That's indeed very similar to the above model. Going from the distribution to the realization corresponds to a step downwards, and moving upwards depicts very nicely the bifurcation one encounters. Since the degrees of freedom rise when moving upwards, but also since we have augmented the above model (replacing a wildcard by a certain distribution) we have to determine the value of p. Due to symmetry that is, because 0 and 1 have exactly the same status, $p = 1/2$ is the most natural choice. Moreover, information theory can substantiate formally that in this case the amount of uncertainty is largest ((Cover und Thomas 2006), p. 16). That is, we are most unsure about the concrete value to occur upon flipping a fair coin.

3.2 Second model

A more typical way to read this situation is, however, just the other way around. 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 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 just 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:

[Illustration 2: Trapezoid with the shorter side at the top]



Altogether the models demonstrate nicely that moving from the general to the particular and back need not involve a major drawback. Far from it, one crucial insight is that information and probability are very much related, and that the inductive step can be treated nicely with the help of probability theory. (That is, a single number or sign may be replaced by a less informative two-point distribution.) 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.

Although important, the notion of more or less general is not crucial. What is crucial is the amount 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, such a step 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 associated with the adjective *inductive*. Combining both directions, a trapezoid, a funnel, a triangle or a tree-like structure may serve as standard models for deductive vs. inductive

moves. 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.

3.3 A general model

There are two ways to approach mathematics: On the one hand, there are concrete objects (certain numbers, vectors, functions, operators etc.), on the other hand, there are assumptions, restrictions, axioms and boundary conditions. Quite obviously, the more or the stronger the conditions, the more specific the situation. The fewer conditions, the more general the situation. Often, objects are defined via very specific conditions, e.g., a square is a right-angled four-sided figure with each side having an equal length. The fewer conditions, the more objects satisfy them. For example, skipping the condition of equal side lengths, the square becomes a less specific rectangle. More generally speaking, conditions and sets of objects satisfying them are reciprocal: the more conditions, the smaller the set of corresponding objects, the fewer conditions, the larger the set of objects.

Consistently, the notion of generality is crucial in (and for) the world of mathematics. Conditions operate "top down", each of them further restricting some situation. 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 A is less general, and any superset C is more general ($A \subseteq B \subseteq C$). Thus our basic paradigm, involving just two sets, is:

[Illustration 3: Two different information bases (more and less general tiers)]

More generally, i.e., 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:

[Illustration 3a: Different information bases (the funnel)]

If the most specific subset consists of a single point one has a triangle-like structure standing on one vertex. The vertex at the bottom corresponds to a uniquely defined object which complies with all boundary conditions:

[Illustration 3b: Different information bases (the triangle)]

The subset relation is also a straightforward way to deal with logical implication, for if $A \subseteq B$, $x \in A$ implies $x \in B$. This corresponds to a simple kind of deduction such as: If I am in Atlanta, I am also in the USA. It is straightforward and not really interesting, since it skips a lot of information about my whereabouts. Unlike the above model (but perhaps closer to usual thinking and discussion) deduction is bottom up: From a particular (small) set to a more general (large) set.

Thus, in this view, induction is top down. This move corresponds to an "insecure" conclusion of the kind "If I am in the USA, I am also in Georgia". Of course, such a conclusion is not watertight. However, it is straightforward to consider it and to ask how much evidence (often ranging from 0 for no evidence to 1 for a certain event) there is. For example, for logical reasons, upon being in the USA it is much more plausible to be in a state west of the Mississippi than to be in California or in Los Angeles. Having population data available one may even assign a probability (frequency) to these events.

Like before, and more interestingly, one may argue the other way around: Often, if a set B has a certain property, then any subset A, and in particular, any element $x \in B$ may also have this property. In this case mathematicians say that the property is *inherited* by smaller substructures. That is, if a property holds in general, it also holds in a more particular situation. A prominent example would be the set of all men, and the property of mortality (i.e., all men are mortal). Consistently, Socrates, being a member of the set, is also mortal. In this sense, i.e., if a property of the whole set B also applies to any subset $A \subseteq B$, inheritance is tantamount to logical implication. Thus deduction is top down.

That's exactly how mathematics works: On the theoretical side one is looking for theorems, i.e., general and strong statements, interesting properties, in particular symmetries, that hold for a large set of objects. On the practical side one applies these theorems to some particular situation, abiding by the assumptions of the general theorem. Thus if I want to calculate the hypotenuse c of a particular right-angled triangle, I may apply Pythagoras' theorem, stating the relation $a^2 + b^2 = c^2$ for any such triangle. Note that there is some tension here: Typically, the larger the set of objects, the more complicated the situation and the more pathologies arise (no nice theorems, or no general theory at all). The smaller the set of objects, the more regular the situation which can be described by elegant theorems. However, these beautiful results can hardly be applied, since they only hold "in a sandbox", i.e., given peculiar circumstances. A typical example is the theory of homomorphic functions, arguably the most elegant of all mathematical theories. However, for the same reason - i.e., the very strong assumptions involved - it is rather difficult to apply.

By symmetry, the interesting direction of induction in this model ascends from set to superset. Since a set has many supersets, there are many different ways to generalize (corresponding to the "funnel"). For example, the natural numbers can be embedded into the large sets of the rational, the real and the complex numbers. They can be interpreted in a geometrical context (e.g., as the length of a line, the surface of a ball, or the volume of a body), an algebraic context (e.g., as coefficients in some equation) or as constant functions, to name but a few.

However, very often a class of objects somewhat more general than a received, rather well-known class of objects offers itself to investigation (that's why the Greeks moved from rational numbers to reals). Or, since concrete objects are associated with a large

set of assumptions, it is rather straightforward to withdraw some of these and study the properties of the larger class of objects thus defined. Such steps of abstraction (i.e., subtracting some property), aiming at a well-behaved larger class of objects, could also be called "conservative induction".

For example, within the rather small set of real numbers $\mathbb{R}$, multiplication is commutative, that is $a \cdot b = b \cdot a$ holds for all $a, b \in \mathbb{R}$. One may embed the reals into the larger set of 2×2 matrices. That is, some real number r is represented by the diagonal matrix $\begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}$. It turns out that for general 2×2 matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the commutative law no longer holds. In other words, the condition that multiplication be commutative needs to be withdrawn, since in this class of objects, $A \cdot B \neq B \cdot A$.

Arguably the most basic "drive" of mathematics consists in generalizing results. More precisely, in everyday work, mathematicians take pains to extend some interesting theorem. Holding the assumptions fixed, they try to strengthen the statement in order to say more about a certain situation; or, given some statement, they try to prove this result with fewer (weaker) assumptions, i.e., in a more general setting. For example, Gauss proved long ago that the sum of independent random variables, each of them having a $B(p)$ distribution, converges to the normal distribution (which is the so called central limit theorem for independent Bernoulli random variables). It took mathematicians many decades to generalize this theorem to independent random variables with quite arbitrary distributions, and to allow some dependencies among the random variables. Perhaps surprisingly, in this sense, the course of mathematics in history has been inductive, from particular examples to general concepts, and from small sets of objects to large spaces.

3.4 Bounded generalizations

So far, induction has not posed insurmountable problems. However, thinking of Hume's problem, many inductive steps do not seem to be harmless but rather highly problematic. Illustration 3 gives a reason why: Induction is like a funnel defined by two sets (a less and a more general one). In the case of a "wild" generalization, i.e., upon skipping many constraints or upon enlarging the class of objects under consideration enormously, the upper end of the funnel is quite far away from its bottom. Then, one may encounter a situation which no longer has "nice" properties. For example, calculus is a rather restricted field, since functions have to obey rather far-reaching conditions (in particular differentiability). Reducing these assumptions to a bare minimum, mathematicians developed the field of general topology. Equivalently, a topology defined on an arbitrary set is a very general mathematical structure. Thus this field needs a myriad of concepts to classify the enormous number of phenomena occurring. The situation resembles a high mountain range, with sharp peaks, enormous chasms, and few accessible paths. Straightforward ideas typically do not work, large detours are the rule, and consequently a famous classic book lists almost 150 (different classes of) "counterexamples in topology" (Steen and Seebach 1995).

Could there be a situation when the upper set, representing the general situation, is "too far away" from the lower set? In the case of the first model, where the distance can be measured by means of the number of yes-no questions, surely not. In the case of mathematics, where it does not make sense to skip all assumptions, one also has,

at least in principle, a bounded step from the particular to the general. As long as we do not leave set theory, the basic assumptions of set theory, i.e., one of its axiom systems, keep the situation at bay. However, when Cantor was looking for a general framework for mathematics, he did not know the axioms. Moreover, he did not keep close to the structures already known, but generalized boldly. In particular, he knew that mathematicians at times venture into the potential infinite. That is, given a finite structure, a theory can be made more elegant by adding an ideal infinite element. For example, any two non-parallel lines in the plain intersect at one point. It would be nice if all lines had this property. Playing around with straight lines it soon becomes obvious that parallel lines should intersect at the ideal point "infinity" which in a sense lies beyond the plain. Another prominent example are convergent series. For example, if you summarise consecutive powers of two, $2^0 = 1, 2^0 + 2^{-1} = 1 + \frac{1}{2} = 3/2, 2^0 + 2^{-1} + 2^{-2} = 1 + \frac{1}{2} + \frac{1}{4} = 7/4$, could there be something like a limit if you added up all potential summands? In other words, can the infinite sum $\sum_{i=0}^{\infty} 2^{-i}$ be defined, and if yes, what is its value?

To cut a long story short, this kind of bold, inductive mathematics, leaving well-known particular structures far behind, went too far. Naive set theory taking on "actual infinity" encountered logical fallacies. Fortunately, these pitfalls could finally be remedied within formalized set theory; more precisely, the axiom of foundation excludes viscously circular structures and infinite continued descents. Similar stories happened when the number systems were gradually extended. The Greeks were shocked when they encountered - mind the name! - irrational numbers. Having finally understood the reals, "imaginary" numbers were added, leading to the complex number system and beyond.

The point is that a sound upper layer first has to be constructed in order to make the inductive step work. That it took so long to reach larger number systems is not a coincidence. The task is intrinsically more difficult than working within a given framework.

There is a similar problem with respect to the lower layer. The more assumptions, the fewer objects satisfy them all. Too many assumptions, and there will be no object left. (For example, there is no number x that is both positive and negative.) In other words, if the lower set disappears, the difference between the upper and the lower set of objects is undefined. Since mathematics is rather oriented towards generalization, this problem is, however, rather academic. (A notable exception is provided by Arrow's impossibility theorem, see Maskin and Sen (2014).)

As long as the distance between the more general and the more specific situation is in some sense bounded, induction is acceptable, and since we are in the formal world, infinity is not really the issue. This being said, it seems to be reasonable and straightforward to replace a finite number of data points by a single smooth function, which is what regression, the main workhorse of econometricians, does. Of course, this step is not unique or deductively determined, but given the data and a reasonable criterion, it becomes a standard optimization task, treatable within a strict mathematical framework. Fisher (1935/1966), p. 4, also follows this strategy and refers to "games of chance" as a typical example. Altogether, he concludes:

> 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.

3.5 Convergence

Suppose the distance between the upper and lower set is bounded, the upper set having more content than the lower one, and thus the inductive direction is bottom up. It is reasonable to call an inductive step acceptable if it is sufficiently small, or when it can be made arbitrarily small in principle. For this basic setting, Saint-Mont (2011) gives a convergence argument based on boundedness, essentially answering Hume:

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) or to add observations (lifting the lower layer upwards). Inserting further layers in between, the inductive gap can be made arbitrarily small. In particular, the charge of circularity (the inductive step being based on an inductive principle, i.e., a further assumption) can be treated in the following way:

Like all further assumptions, an inductive principle helps to bridge the gap in lowering the upper layer. (The general law becomes a bit more special.) If this suffices to shrink the gap below the criterion defining an acceptable inductive step, we are done. Otherwise, one iterates this procedure, i.e., one evokes a second inductive principle, supplementing the first. If this sufficiently closes the gap, we have succeeded again. Otherwise, we need to evoke further inductive principles, bringing up Hume's charge of infinite regress.

However, an infinite regress in this case is harmless, since the consecutive gaps can be understood as a convergent series. In other words, the additional assumptions that have to be evoked to close the inductive gap tend to become *weaker*, finally making the resulting gap arbitrarily small, i.e., essentially closing the gap. 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 quite a bit of mathematical subtlety to find a theoretically satisfying answer.[2]

In the case of an unbounded gap that's different, the sequence of successive gaps need not disappear, and therefore the property of boundedness is crucial.

The alternative strategy is to add observations. For example, 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

[2]Interestingly enough, Rescher (1980), pp. 208-209 also uses an iterative sequence of inductive arguments, quite similar to Illustration 3. Although he mentions Achill and the tortoise explicitly, he fails to realize that a precise notion of convergence has dissolved this paradox. For another concrete example see section 3.6 which discusses the classic that the sun is supposed to rise tomorrow since it has always risen in the past.

to the persons not yet investigated. Now, since we are dealing with the problem in a formal way, statistics has no qualms about dealing with infinite populations, and the upshot of sampling theory is that even in this case a rather small, but carefully chosen subset suffices to get "close" to properties (i.e., parameters) of the population. Not surprisingly, probability theory plays a major role here and the various philosophical schools of statistics disagree on how to apply it. Although this case looks like enumerative induction à la Bacon (which is not convincing, see section 5.1.1 below), notice that we are operating in a strictly formal framework here.

Thus the assumptions built into this framework guarantee reasonable behaviour, in particular, convergence of the sample (and its properties) towards the larger population. It may be mentioned, however, that one of the most prominent statisticians of the 20th century (de Finetti 1937), realized that any finite sample contains too little information to pass to some limit without hesitation. In particular there has been some discussion about 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, in general, mathematicians need and indeed just assume countable additivity: $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. For more on this topic see section 5.5.

Although convergence is (or at least seems to be) a straightforward concept, a few common caveats should be mentioned. First, convergence need not be monotonous. It is true that a bounded monotonically increasing (or decreasing) sequence must converge. However, that is just a sufficient, but not a necessary criterion for convergence. That is, there are also non-monotonic sequences that converge. So, with respect to induction, if a sequence is non-monotonic, this does not suffice to render a convergence argument invalid. Second, convergence is a property of infinite sequences, and only the tail of the sequence is relevant. This means that any finite subsequence is irrelevant. In particular, the beginning $x_1, \ldots, x_n$ of an arbitrary sequence does not give any information about convergence of the whole sequence $(x_i)_{i \in \mathbb{N}}$. In this sense, highly regular sequences typically studied by mathematics, such as $(2^{-i})_{i \in \mathbb{N}}$ or $(1/i)_{i \in \mathbb{N}}$, are rather misleading. Third, convergence may be arbitrarily slow. Thus counterexamples that implicitly assume a certain rate of convergence are also missing the point. In particular, if convergence is slow, any initial but finite part of the sequence (i.e., $x_1, \ldots, x_n$) gives virtually no information on the sequence's limit.

3.6 The sun will rise tomorrow, will it?

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. But is this really so? 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, let S_i be 1 if the sun rises on day i and 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 $n \to \infty$.

It may be added that Laplace (1812), who gave a first 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 should be monotonically decreasing. 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). Already Laplace noted that the probability becomes (much) larger, if further information about our solar system is taken into account.

The above may look as if 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 just 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.

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's not true. 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 $\Delta(k) = I(k+1)/I(k) = \frac{k+1}{n}/\frac{k}{n} = \frac{k+1}{k} = 1+1/k$ which is a decreasing function in 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 (Cover und Thomas 2006: chap. 3). Rissanen (2007: 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.[3]

3.7 Nonviscous recursion

Apart from such considerations, there is another straightforward way to look at this example in order to show that the alleged "vicious circularity" does not hold water. 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, recursive functions (loops) demonstrate that this is not so:

$n!$, called "n factorial" is defined as the product of the first n natural numbers, that is $n! = 1 \cdot 2 \cdots n$. Now consider the following program FACTORIAL[n]:

> IF $n = 1$ THEN $n! = 1$
>
> ELSE $n! = n \cdot (n-1)!$

At first sight, this looks viciously circular, since "factorial is explained by factorial" (the factorial function is used to calculate itself). However, 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 example,

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

The point is that on closer inspection the factorial function has an argument (called a parameter in IT-jargon). Every time the program FACTORIAL[n] 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. Due to boundedness, this strategy leads to a well defined algorithm, calculating the desired result.

As a consequence, in general, using $I(n)$ in order to get to $I(n+1)$ need not be viciously circular. If, the smaller n, the weaker the assumption $I(n)$, the fewer content $I(n)$ has, there is no (vicious) circularity. 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 tautologies or deductive arguments in general, one would be begging the question. However, since we are dealing with induction, i.e., inevitable inductive gaps, a (finite) sequence of assumptions becoming weaker and weaker is nothing but a certain kind of convergence argument, discussed in subsection 3.5. (Of course, if the gap $I(n+1) - I(n)$ is large, the inductive step need not be convincing, but that is a different question.)

[3] 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.

3.8 Information theory

The very first model considered belonged to information theory. Due to its finite nature, it 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* 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. The crucial technical steps are as follows:

(i) **Terminology**. $l(\mathbf{x})$ is called the length string $\mathbf{x}$ (measured in bits). Programs $\mathbf{p}$, serving as input are also called models for $\mathbf{x}$ or hypotheses consistent with $\mathbf{x}$. In a sense, $\mathbf{p}$ contains all the information or regularity in $\mathbf{x}$, and since it should be smaller than $\mathbf{x}$, it can also be thought of as a kind of compressed version of $\mathbf{x}$. Thus, induction, here, amounts to modelling data in a concise way, its guise being the art and science of data compression.

(ii) **Kolmogorov complexity**. Let $\mathcal{U}$ be a universal computer (e.g., a Turing machine) and let $\mathcal{U}(\mathbf{p})$ denote the output of the computer $\mathcal{U}$ when presented with the program $\mathbf{p}$. Then the Kolmogorov complexity $K_{\mathcal{U}}(\mathbf{x})$ of a string $\mathbf{x}$ with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(\mathbf{x}) = \min_{\mathbf{p}:\, \mathcal{U}(\mathbf{p})=\mathbf{x}} l(\mathbf{p})$$

In other words: Kolmogorov complexity of a string $\mathbf{x}$ with respect to computer $\mathcal{U}$ is the length of the shortest computer program that, when given to the computer $\mathcal{U}$ as input, results in the output $\mathbf{x}$. As is to be expected, $K_{\mathcal{U}}(\mathbf{x})$ is not computable. Moreover, it is crucial that $K_{\mathcal{U}}(\mathbf{x})$ can be interpreted as the complexity $K(\mathbf{x})$ of $\mathbf{x}$, i.e., it is essentially independent of the particular computer used.

(iii) **Universal probability**. Now, probability comes in in a natural way: The probability that a program $\mathbf{p}$ occurs randomly, i.e., by means of flipping a fair coin sequentially, is defined as $2^{-l(\mathbf{p})}$. (Note that thus shorter programs are considered much more probable than longer ones.) The universal probability of a string $\mathbf{x}$ is the probability that some program randomly occurring as a sequence of fair coin flips will print out the string $\mathbf{x}$,

$$P_{\mathcal{U}}(\mathbf{x}) = \sum_{\mathbf{p}:\, \mathcal{U}(\mathbf{p})=\mathbf{x}} 2^{-l(\mathbf{p})}$$

(iv) **The result**. "The overarching principles put together by Solomonoff (1964) are: (Universal) Turing machines (to compute, quantify and assign codes to all quantities of interest), Kolmogorov complexity (to define what simplicity/complexity

means), Epicurus' principle of multiple explanations (keep all explanations consistent with the data), Ockham's razor (choose the simplest model consistent with the data)."[4] Li and Vitányi (2008), p. 347, elaborate: "Essentially, combining the ideas of Epicurus, Ockham, Bayes, and modern computability theory, Solomonoff has successfully invented a perfect theory of induction. It incorporates Epicurus's multiple explanations idea, since no hypothesis that is still consistent with the data will be eliminated. It incorporates Ockham's simplest explanation idea since the hypotheses with low Kolmogorov complexity are more probable. The inductive reasoning is performed by means of the mathematically sound rule of Bayes." "In the end, we choose the simplest explanation that is consistent with the data observed" (Cover und Thomas (2006), p. 490).[5]

4 Empirical framework

Empirical science is based on data, in particular data stemming from (systematic) observations. Abundant as they may seem, they are always bound to a specific situation, e.g., an experiment that was conducted at some time in a particular place. Thus, in this sense, the set of all data is rather small.

The other side of the story is theory. General concepts, rules, structures and patterns that govern the phenomena of some field. That is, they try to explain how the data was created. One need not evoke some specific model of explanation (e.g., the deductive-nomological) in order to agree on the fact that the theoretical tier may be (very) general, whereas the observational tier is (much more) specific: There is an enormous gap between the colour of a particular swan I see, and a statement like "all swans are white".

4.1 Mind the - bounded - gap

Given the last example, the gap seems to be unbounded, and thus any inductive effort doomed. Perhaps that's why an eminent philosopher like Popper abandoned induction altogether, and a thoughtful scientist like Jeffreys (1973), pp. 14, 58, knowing that he was up against Hume at this point, tried to make the gap as small as possible:

> What [a scientist] actually does is to state his laws in the most general and simple form possible, and modify them as later experience indicates [...] But the most that has ever been verified is that some general laws have had no exceptions hitherto (and it is very difficult to find such laws). It is not verified that any accepted general law will always hold in the future.

Our reply is different: We acknowledge that data are very specific, in particular if they stem from a carefully designed and thus rather "artificial" situation, such as a highly sophisticated experiment in contemporary physics. We also acknowledge that laws can be extremely general, just think of physics' most esteemed findings, like the

[4] See Hutter (2007), p. 38, order of arguments has been reversed.

[5] Almost equivalently, Kemeny (1953), p. 397, wrote much earlier: "Select the simplest hypothesis compatible with the observed values."

conservation of energy. Nevertheless, we claim that the gap between them is bounded, and thus - in principle - good-natured.

In a nutshell, we are dealing with matters of fact here. That is, a quantifier like "all" cannot be understood in an infinite way. If we read the quantifier in an infinite way, we would be confounding relations of ideas with matters of fact, a principled distinction drawn by Hume. General laws of nature must be understood in a different way, at the very least one must distinguish thoroughly between their mathematical-formal aspect and their empirical content.[6] Within any formal treatment, the idea of assumptions restricting generality is crucial; in complete analogy, boundary conditions lead the way here. More precisely, any experiment is conducted in a very restricted situation, given severe boundary conditions. Nevertheless, science aims at generalizing any such result beyond the laboratory.

Since in all sciences there are two kinds of factors - relevant and irrelevant -, it is a common empirical strategy to deliberately modify the situation in order to isolate the relevant factors. Skipping all irrelevant conditions, the situation also becomes more and more general. Physics, having the most general laws (and which may thus serve as a model here), succeeds in isolating the very few factors that are really important for a certain phenomenon. Fortunately, these and their relations may often be summarized and presented in a concise theory, perhaps a single formula.

What is left besides the very few relevant factors - depending on the subject area - are conditions that always hold. If some law always holds, this does *not* mean that it holds unconditionally. More precisely speaking, it holds under conditions, that are always fulfilled in our universe. In the world we live in, there is no setting or place that is void of conditions. All that can be reached on the most general empirical layer are structures that are omnipresent, in existence everywhere and anytime. (There is no perfect vacuum either. Rather, vacuum in its "purest" form has a certain structure to it.)

In other words, generalizing to the very limit in the empirical world does not lead "nowhere". Rather, it offers a definition of a law of nature: Such a law is a general principle that, *given the boundary conditions of our universe*, holds without any further particular restrictions. Proceeding in the deductive direction, it is absolutely no coincidence, but top-down necessity (i.e., Noether's theorem) that the basic symmetries of space and time imply the (al)most fundamental laws of nature. Given certain invariances that always hold, i.e., everytime and everywhere, imprinted in the structure of space and time, the major conservation laws are their logical implications. For example, invariance with respect to time translation yields the law of energy conservation.

These laws are themselves boundary conditions, standards that have to be met by every particular object and phenomenon. Thus there is no perpetuum mobile, and no object may travel faster than the speed of light. Instead, every process and every object existing in time and space has to comply with basic physical laws. In this sense, chemistry is nothing but physics within particular marginal conditions, biology is constrained chemistry, and psychological processes must be compatible with physiological prerequisites.

[6] "...as far as the propositions of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality." Einstein (1954), p. 233.

4.2 Hidden structures and mechanisms

Bridging the bounded but nevertheless enormous abyss between the general and the particular is far from easy. And it is not just the width that is crucial: Since physical laws have a rather simple structure and since collecting informative data is rather easy in this area, physics succeeded long ago and time and again in creating good theories. In the social sciences, this is different: There is an enormous number of factors that are difficult to access and to separate. Moreover, they influence one another and the obtainable observations in a confusing, often very dynamic way. Thus erecting a small and stable structure on sand or even quicksand can be much harder than building a skyscraper on solid ground (Meehl 1990).

Fortunately, very often the theoretical layer is *not* an extremely general law, hidden behind clouds of abstraction. It is rather typical that the abstract layer consists of an interesting, general structure or mechanism that needs to be uncovered within a bunch of irrelevancies. There is much data and a rather rich context, but it is difficult to detect the signal within a lot of noise.

Here again, luckily, substantially different background structures and latent mechanisms imply completely different kinds of data and, because of substantial prior knowledge, there are often just a few plausible mechanisms. Therefore search can be *restricted* to quite a limited number of candidates, making experimentae cruses possible. At times, the core problem consists in finding a single candidate that is compatible with general theory and crucial data.

Consider diseases: On the theoretical side a pathological process has to comply with all that is known about the body, its anatomy and physiology. On the practical side, symptoms of certain diseases are quite typical or even unique. Of course, diagnosis is a tricky business, but physicians have come a long way in uncovering causal pathways. In this endeavour, it is quite an advantage that the true mechanism can be approached top down, i.e., via restrictions stemming from theory, and bottom up, i.e., by means of interpreting existing data. Typically, knowledge accumulates at the bedside and in the laboratory, with the result that successful cures go hand in hand with a deeper theoretical understanding of relevant factors and their interactions. Contemporary translational medicine's agenda is to find effective therapies that span the chasm between "bench and bedside".

Of course, in principle, one cannot verify a latent variable, structure or mechanism once and for all. Since no scientific theory can be proved, there always remains some fundamental doubt. But one should not exaggerate this point: Given a natural system, such a system has a certain structure to it and a finite number of processes that are going on. The facts can be very complicated, technical difficulties may seem almost insurmountable, and completely new theoretical concepts might be necessary. Thus, it may take a long time to solve a certain puzzle. However, once we have found the solution to a (bounded) problem, in particular, when we are able to give a detailed account of what is "really" going on, such a discovery is there to stay. Cell theory, the periodic system of elements and many other major insights are (almost) final and absolute.

4.3 Models as mediators

In general, models mediate between data and the conceptual level (Giere 1999, Morgan and Morrison 1999). An ideal model connects the two layers in an elegant way, since it captures all crucial factors, and combines them in an (approximately) right way. Thus explicating all important variables and their functional dependencies, one gains some understanding into what is "really" going on. When fed with realistic data, such a model fits well to reality, in the best case it is able to forecast with high precision (Lehmann 1990, Cox 1986, 1990, 1995, 2000). In his 1990 contribution, p. 169, the latter author elaborates:

> In many ways, the most appealing models are those that connect directly with subject-matter considerations [...] These models aim to explain what is observed in terms of processes (mechanisms), usually via quantities that are not directly observed and some theoretical notions as to how the system under study 'works'.

To this end, one may either start with theory or with data, and each starting point comes with a different emphasis on simplicity vs. fit. In the latter case, statistical efforts to condense an abundance of data and a long line of factors into a concise model have been named data-driven analysis, data mining, hypotheses generation, knowledge discovery, machine learning, and, quite simply, modeling. Since taking into account more factors typically improves fit, there is a tendency to build empirical models "as big as an elephant" (Savage). Today, with an abundance of computer power able to deal with gigabytes of data and thousands of potentially relevant variables at once, there are impressive applications in the fields of automatic (deep) learning, neuronal networks, and artificial intelligence (robotics).

In the extreme, however, one is reminded of Wiener's famous aphorism that "the best material model of a cat is another, or preferably the same, cat." In other words, there is the very real danger of overfitting, i.e., some model cannot be transferred to a similar set of data. Freedman (2010) pointed out that very often data do not speak for themselves - even if "torturing them until they confess" - and using methods from the shelf (e.g., multiple regression) has proved to be rather disappointing. The bottom line is that without theoretical insight, one easily drowns in data, resulting in bad models, and confusion - but not scientific breakthroughs.

Starting at the other end of the chasm, i.e., driven by theory and a conceptional underpinning, there are so-called "explanatory models". These endeavours rather favour "minimal models", only taking into account a few crucial features. This approach works well if there is a precise and adequate theory to build on, covering the problem in question. Thus in physics and other natural sciences, (applied) modelling has become a standard. In the best case, given a well-defined situation and sound background knowledge, a simple model catches the essence of some real-life phenomenon, and predictions boil down to mathematical calculations.

However, in many fields, e.g., economics, theory is less reliable. Thus the foundations one is building on are shaky and it is also inevitable to idealize matters to a large extent. The result of such endeavours are so-called "toy models". Since their intrinsic structure is rather plain, it is possible to gain some theoretical insight, i.e., to understand some interesting phenomenon better. However, oversimplifying matters may also lead one astray easily, and, when applied to real data, such a model may collapse at once.

Forecasts may be far away from reality, in particular in "extreme situations" when the need for theoretical guidance would be largest. For example, hardly anyone predicted the major economic crisis of 2008, and no model so far has been able to tell practitioners how to overcome it effectively.

Models are best if they are able to join constructions started on each side of the gap. Considering the example (i.e., history) of weather and climate prediction, Edwards (2010) shows what a tremendous amount of effort may be necessary to make ends meet. On the one hand, basic physical theory and meteorological processes have had to be translated into calculable numerical problems. On the other hand, worldwide, piles of data need to be collected and processed in real time. In nuce, these contributions materialize in a multitude of specialized modules, being combined in immensely large models. Finally, the results of various models have to be combined into a global, but at the same time also very detailed picture. Needless to say these efforts require an enormous amount of IT-equipment and seemingly endless numbers of calculations.

4.4 Successful experimentation

A bounded formal gap can be made smaller by adding constraints (assumptions) "top down" or by adding objects/data "bottom up". Analogously, there should be general strategies to bridge an empirical gap.

The way "top down", traditionally called "deductive nomological", seems to be rather straightforward. In the easiest case it consists of a formula that can be applied to a particular case. One just has to replace general variables by particular numbers which may amount to filling in a form. More generally speaking, there is some theory that can be customized to the problem at hand. Rather typically, however, some theory is a less powerful constraint; it helps in finding the way but does not clear all obstacles.

The way "bottom up" is tedious. Since it is very difficult to extract valuable information from data per se, it is a good idea to collect data within a rather strict framework. Systematic experimentation, i.e., augmenting data with additional structure, putting data in a strong context, has proved to be a reliable way to new knowledge. In particular, any controlled experiment tries to exclude unwelcome "nuisance" factors, thus making alternative explanations unlikely. Replication is a related, very powerful technique: First, it can be used to establish the reality of some phenomenon, i.e., that some finding wasn't a random quirk. Second, by systematically changing conditions, one is able to sort out which factors are important and which are not. Finally, a "constructive" replication (Lykken 1968) is only possible, if relevant influences are known, nuisance factors can be held at bay, and if theoretical concepts are clear enough.

Systematic experimentation in the natural sciences goes back at least to Galileo, and Fisher (1935/1966) extended "experimental design" to areas plagued with a lot of variance (many unknown nuisance factors). During the last decades, his ideas have been extended to a rather comprehensive theory of causal inference (e.g., Fisher (2003), Shadish et al. (2002), Pearl (2009), Freedman (2010)). Nowadays, criteria such as objectivity, reliability and validity are are commonplace, and, in particular, a study is called "externally valid", if explicit precautions have been taken to justify the generalisation of the study's result to a similar situation.

A forerunner to Fisher's ideas were J.S. Mill's methods of generalization (see Skyrms (2000), chap. 5, for a contemporary exposition). Most importantly, his method of differences says that

> If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occuring only in the former: the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon. (Mill 1843: 225)

A typical contemporary experiment in the social and medical sciences follows this logic: Two similar groups are compared. One of them is given a certain treatment T, whereas the other group serves as a (local) control C, receiving no treatment. Given this design, because of Mill's argument, an effect observed in the end must be due to the difference between T and C, i.e., the particular treatment.

Let us finally mention that statisticians Cornfield und Tukey (1956) emphasized long ago that a generalisation must be based on formal as well as on substantial considerations:

> In almost any practical situation where analytical statistics is applied, the inference from the observations to the real conclusion has two parts, only the first of which is statistical. A genetic experiment on *Drosophila* will usually involve flies of a certain race of a certain species. The statistically based conclusions cannot extend beyond this race, yet the geneticist will usually, often wisely, extend the conclusion to (a) the whole species, (b) all *Drosophila*, or (c) a larger group of insects. This wider extension may be implicit or explicit, but it is almost always present. It we take the simile of the bridge crossing the river by way of an island, there is a statistical span from the near bank to the island, and a subject-matter span from the island to the far bank. Both are important.

Of course, these considerations only make sense if the far bank isn't infinitely far away. Only a bounded stream may thus be bridged. Interestingly enough, if a choice between formal and subject-matter considerations has to be made, the former authors strongly advise taking both aspects into account:

> By modifying the observation program and the corresponding analysis of the data, the island may be moved nearer of farther from the distant bank, and the statistical span may be made stronger or weaker. In doing this it is easy to forget the second span, which usually can only be strengthened by improving the science or art on which it depends. Yet an unbalanced understanding of, and choice among, the statistical possibilities requires constant attention to the second span. It may often be worth while to move the island nearer to the distant bank, at the cost of weakening the statistical span - particularly when the subject-matter span is weak.

Bounded generalisations are not just commonplace in science; successful, as they typically are, they are a major part of science's glory. Since the philosophy of science (i.e., Popper, Kuhn, Lakatos, Feyerabend and their successors) has become very sceptical -

for purely theoretical if not fundamental reasons one might add - there seems to be a renewed interest in these matters among scientists. Some like Jaynes (2003), Li and Vitányi (2008) address induction, but there are also up-to-date books on the scientific method in general, e.g., Gauch (2012), and Bookstein (2014). Typically, these blend subject-matter and formal arguments, and the methods they discuss can be understood as specific empirical answers to Hume's problem.

4.5 Precision makes a difference

Fisher (1935/1966), p. 102, is surely right when he speaks out against highly standardised experiments - since they narrow the "inductive basis" of our conclusions. However, looking at physics, collecting more data in a variety of settings, although important, does not really seem to hit the nail on the head. Physicists rather emphasize the importance of precision, in particular when it comes to measurements. Instead of breadth they count on depth: They go to extremes and check if crucial theoretical values coincide - precisely - with their empirically measured counterparts. Intuition agrees with this strategy: It is much more difficult to pass a single but tough test than to pass ten easy tests.

Here's a quantitative argument supporting this point of view: Suppose a single measurement has standard deviation σ. Given n such measurements, the precision, i.e., standard deviation, achievable via averaging is $\sigma/\sqrt{n}$. Increasing the precision of the single measurement by a factor of 10, i.e., by one decimal place, increases the precision of the average to $\sigma/(10 \cdot \sqrt{n})$.

Now, in order to achieve a certain precision ϵ via measurements with standard deviation σ, how many observations are necessary in the less and the more precise case, respectively? Straightforwardly,

$$\epsilon = \frac{\sigma}{\sqrt{n}} \Leftrightarrow \sqrt{n} = \frac{\sigma}{\epsilon} \Leftrightarrow n = \frac{\sigma^2}{\epsilon^2} \quad \text{and} \tag{1}$$

$$\epsilon = \frac{\sigma}{10 \cdot \sqrt{n}} \Leftrightarrow \sqrt{n} = \frac{\sigma}{10\epsilon} \Leftrightarrow n = \frac{\sigma^2}{100\epsilon^2} \quad \text{if precision is tenfold.} \tag{2}$$

As an example, suppose $\sigma = 1$. In order to achieve a final precision of $\epsilon = 1/20$, according to equation (1), one needs $n = 400$ observations. However, measuring just one digit more precisely, according to equation (2), the number of observations needed is just $n = 4$. In general, improving the precision of a single measurement by a factor of 10 (one more valid digit), means reducing the number of observations by a factor of 10^2. k more valid digits mean that a single observation obtained in such a precise experiment is worth 10^{2k} observations in the original imprecise experiment. In this sense, measurement precision exponentially broadens the empirical basis of some general statement.[7]

However, and perhaps surprisingly, there is more to numerical precision than just quantitative considerations. In the present context we conceive of a theory as an edifice supported by a foundation of empirical data, more precisely, an abstract layer resting on (a layer of) specific observations. Few or imprecise data are like unreliable building

[7]Note that the whole argument is based on a finite deviation. If $\sigma = \infty$, however, adding observations is futile, if $\sigma(\bar{X})$ is also unbounded. Thus, in this case, using Hume's words, "a hundred or a thousand instances are ... no better than one." A popular statistical model for this situation is the Cauchy distribution.

blocks: They support huts or houses, but not larger structures. The same with verbal arguments: They are good as far as they go, but there is a natural limit as to what can be achieved with their help. In order to reach the abstract heights of excellent scientific theories, in particular physics, one needs more resilient construction material. This is provided by highly precise measurements of the theory's crucial ingredients, but also by mathematical structures and logically stringent arguments holding the numbers together. In other words, with improved (quantitative) precision goes a completely different quality of theory. In order to get to this level, chance observations and imprecise verbal arguments won't do. It's quantitative precision that causes qualitative change, paving the way to logically sound explanations and deeper understanding. This is not to say that imprecise verbal methods are impotent, they just do not suffice to build theoretical skyscrapers.

By the same token, purely verbal arguments should be regarded with suspicion, since many of them contain flaws that only become obvious upon their formalization (e.g., the tale of Achill and the turtle). In the worst case, problems are not solvable with words alone, but their continued discussion leads one astray. Consider Howson (2000), pp. 14:

> Entirely simple and informal, Hume's argument is one of the most robust, if not the most robust, in the history of philosophy.

Perhaps no constructive answer is known *because of* the verbal imprecision of the argument. (You need induction to rely on induction, etc.) In other words, "simple and informal" had better be replaced by "owing to verbal *fuzziness*, the argument has been very robust." A closer inspection of inductive steps, however, reveals that formal models as well as empirical strategies have the strength to cope with it. In a sense, they are able to "uncover" Hume's evasive action.

4.6 Making it work

All the examples given in this chapter have been successful in the sense that the two tiers could be connected in a constructive way: Laws of nature that govern the behaviour of phenomena in any given situation, formal and subject-matter arguments bridging the (bounded) gap between facts and theory, in particular by means of experimental design, precision, replication and modeling. These fertile interactions - i.e., the modern scientific method - have brought about reliable knowledge and a constant stream of progress: The better we have thus understood nature, the more we have been able to build non-trivial technology.

It is important to note that even if the connection is not perfect, it is still working. In other words, if there is some imbalance (e.g., too much theorizing, or too many confusing facts), progress may typically slow down but need not vanish. In particular, biologists and chemists of the 16th to the 19th century did quite well without a basic theory. Nevertheless, it is important to understand when the basic setting tends to become infertile. First of all, there may be no theoretical tier. But without theory, all one gets is a multitude of separate examples. Second, conceptualization may be close to empirical facts, leading to many theories, all just applicable to a rather small range of phenomena. Things seem to work quite well if theory and practice are on a par. Yet if experience is an appendix to theory, i.e., if data is massively theory-laden, then

the framework tends to become an ideology, since then information outside received thinking is rather ignored than embraced.

Theoretical efforts are also rather fruitless if the tiers do not interact sufficiently, and this might be the most important lesson to learn at this point. For example, if the facts are not given the chance to disprove hypotheses, the latter will proliferate, in the end producing much speculation but no fact-based worldview. This tendency is amplified upon focusing on hardly observable constructs and scholarly disputation, but also upon neglecting mathematics - since it is difficult to falsify vague statements. It is no coincidence that in "dark" mediaeval times, before the dawn of modern science, dogma rather disdained empirical facts, and, despite bookishness, progress was very slow or even non-existent. With the wisdom of hindsight that is quite obvious: Medicine could not be improved as long as studying the human body was taboo, precise mathematical arguments were the exception since attention focused on the afterworld, and the unassailable status of the Ptolomaic system almost made Galileo a martyr. Today, theoretical physics - but also parts of the humanities, sociology and philosophy - have a touch of aloofness, since string theories and their ilk can be made compatible with almost any observation. Compare this to a crisp experimentum crucis: Discussion could be narrowed down to just two hypotheses, and there is data to decide among them.

In a nutshell, modern methods work since they are empirical-experimental as well as conceptual-theoretical. It is the ongoing interplay between facts and conceptualization that turns out to be successful. In this view, the essence of fruitful research efforts is an *iterative* deductive-inductive process that aims at deepening theoretical understanding and at broadening the scope of applications (Saint-Mont 2011).

5 Earlier attempts

So far, we have elaborated formal models and empirical strategies. Next we study received attempts in order to finally give an encompassing answer.

5.1 Bacon: Unbounded generalizations

"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." (Groarke (2009), pp. 80, 87, 79, my emphasis)

5.1.1 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 $\lim x_i = x$.

There are two satisfactory ways to treat the problem: Empirically, it does not make sense to speak of a limit $\lim_{n\to\infty}$, even if n is very large. If the information gathered upon observing the first k observations does not suffice, one may in principle continue with the observations until the whole finite sequence $x_1, \ldots, x_n$ has been observed. Formally, given just a finite sequence, the infinite limit is, of course, arbitrary. Therefore the concept of mathematical convergence is defined just the other way around: Given an infinite sequence, any finite number of elements is irrelevant (e.g., the first n). A sequence $(x_i)_{i\in\mathbb{N}}$ converges towards x if "almost all" (all except a finite set of elements) x_i 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 $x_1, \ldots, x_n$.[8] 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 in order to get a reasonable result. One of these additional rules is Ockham's razor, for a more thorough treatment see section 3.8.

The formal and the empirical part of the question "Will the sun rise tomorrow?" should not be merged carelessly, since then, badly defined problems, having no satisfactory answer, may easily occur. In particular, "uniformity of nature" is exactly such an opaque blend of a formal (uniformity) and an empirical aspect (nature). Formally, the problem can be treated in a satisfactory way (in particular see sections 3.6 and 3.7). Empirically, finiteness makes the problem benign.

Frequentist theory, if it is trying to define probability by means of limits of empirical frequencies, is another 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.[9] 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 will give a good approximation of the number π, however, physics is not able to prove a single mathematical fact about π. Instead, one must define a circle as a certain *relation of ideas*, and one needs to "toss a coin" in a theoretical framework. Of course, empirical problems motivate the definition of an appropriate formal framework, yet they do not have any say when it comes to proof. For more on this matter see section 5.5.

[8] For more details see sections 3.2 and 3.6.

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

Of course, the natural sciences combine both aspects, formal and empirical. However, in doing so, they do not fall prey to the siren calls, asking them to "directly" extend concrete empirical facts to a mathematical sound theory, which is a futile attempt. Moreover, they take pains to avoid any circularity. For example, statisticians use the information $I(n)$ gathered so far to build a model, and use this model for a prognosis (day $n+1$, say). Within the model, in the theoretical world, we are free to do whatever we like, e.g calculate a probability that the sun will rise tomorrow. Trusting this calculation in the real world is a different thing, however. Since the model $M(n)$ has not been evaluated on observations, statisticians would not trust the prognosis. Instead, if no other information is available (i.e., from other sources than just the data at hand), they would use just a part of the observations to build the model, e.g., observations $x_1, \ldots, x_k$, and then evaluate their data-driven model $M(k)$ on the rest of the available data (i.e., observations $x_{k+1}, \ldots, x_n$). A typical choice is $k \approx n/2$. Thus they avoid viscous circularity and have more reason to believe their prognosis for day $n+1$. However, that's all that can be done with the information in the data. Since $I(n) < I(n+1)$, an inductive gap is inevitable. More information available on day n would make the gap smaller, but not eliminate it.

5.1.2 Eliminative induction

Another classical approach, preceding Hume, is *eliminative induction*. Quite similar are Peirce's method of *abduction*, and contemporary *inference to the best explanation* (Lipton 2004). Given a number of hypotheses on the (more) abstract layer and a number of observations on the (more) concrete layer, the observations help to eliminate hypotheses. In detective stories, with a finite number of suspects (hypotheses), this works fine. The same with an experimentum crucis that collects data in order to decide between two rival hypotheses. The real problem, again, is 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's 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 models (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.

A general answer to the positive is given by statistics. Given a finite number of (substantially different) hypotheses, one of which produces the data, a sufficiently large number of observations reveals the true hypotheses with any amount of certainty. In the case of an infinite number of hypotheses, the theory of estimation is able to narrow down the hypotheses considerably. For example, if in successive throws of a coin $P(X = 1) = \theta$ (the probability of seeing Heads, say) is unknown, one may nevertheless give probability statements about the unknown parameter becoming better and better with increasing n. All θ in the open interval $(0, 1)$ may remain possible, but most of the probability will focus on a small interval (a, b) about the true parameter θ_0.[10]

[10] At least in a Bayesian framework. In a Frequentist framework, one would have to say that the "confidence" in the statement $\theta \in (a, b)$ increases, which is weaker. For more details see section 5.5

Of course, the hypotheses need to have some connection with the data. By observing ravens it is impossible to decide whether swans are white. Therefore, in statistics, the abstract level is modelled by a random variable X having a certain distribution (e.g., tossing a coin once; $X \sim B(\theta)$), and $X = x$ is a certain observation (e.g., "Heads"). In this situation, the theorems of mathematical stochastics guarantee that the information gathered on the observational level converges towards the upper level. For example, if $\hat{\theta} = \theta(x_1, \ldots, x_n)$ is a reasonable estimator of θ, we have $\hat{\theta} \to \theta$ in some sense (convergence in probability by the law of large numbers). One of the most general results of this kind is the main theorem of statistics, i.e., that the empirical distribution function $\hat{F} = F(x_1, \ldots, x_n)$ converges towards the true distribution function F of X in a strong sense. In other words: there is a bounded gap that can be leaped over by lifting the lower level (in this view tantamount to more observations).

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 straightforward 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). In more detail, Rissanen (2007), p. 104, writes (emphasis in the original):

> It seems that the real issue in hypothesis testing is to be able to measure how well models fitted to the data are separated. In case of just two models, the problem amounts to calculating the two error probabilities and determining the decision boundary for which the sum of the error probabilities is minimized - i.e., the Neyman-Pearson lemma. The difficult case is when we have a parametric class of models [...] The central problem then becomes how to partition the parameter space into at most a countable number of equivalence classes such that any two adjacent models can be *optimally* distinguished from a given amount of data in a measure that is intuitively acceptable and can also be formally justified.

It may be mentioned in passing that Neyman's and Pearson's famous lemma is by no means the only way to treat the (easy) problem of distinguishing between just two hypotheses subject to a growing number of observations. Closer to the discussion here is theoretical work calculating the probability of misleading evidence, i.e., the probability that data point toward the wrong hypothesis. It turns out that this probability decreases very fast (Royall 2000), i.e., typically, data do not cheat but indicate the true hypothesis.

5.2 Coping with circularity

5.2.1 Black et al.: Self-supporting induction

Black (1958) claimed to have found "self-supporting" inductive arguments. To this end, he notices that "nothing would be accomplished by any argument that needed to assume the reliability of an inductive rule in order to establish that rule's reliability" (p.

718). Therefore, he distinguishes between first-, second- and higher-order arguments, and says "So long as the rule by which the second-order inference is governed differs from the rule whose reliability is to be affirmed, there will be no appearance of circularity" (p. 719). Here is Black's main idea:

> The function of higher-order arguments in the tangled web of inductive method is to permit us to progress from relatively imprecise and uncritical methods to methods whose degrees of reliability and limits of applicability have themselves been checked by inductive investigation. It is in this way that inductive method becomes self-regulating and, if all goes well, self-supporting.

Rescher (1980) can be understood as an elaboration of this approach. His starting point is the observation that there are "epistemic gaps" making "inductive leaps" necessary. Like this contribution, he emphasizes the concept of information:

> The enthymematic model of induction as a question-answering process of truth-estimation casts induction in the role of a gapfilling technique - a method for securing answers to our questions in situations of imperfect information. Its work is genuinely ampliative rather than merely inferential: it does not lie in unravelling the inner ramifications of a preexisting state of informational affairs, but in bringing about a new state through augmenting or supplementing the information at our disposal ... It involves an enthymematic supplementation of given data that is not an inferential step capable of validation by any logical or quasi-logical means, but rather is the product of a *decision* - a decision to bridge over an epistemic gap by a certain data-transcending "act of acceptance."(Cf. Rescher (1980), pp. 58-59, emphasis in the original.)

In order to justify induction, he then follows the Peircean strategy, i.e., he starts with some noninductive "initial justification" (chap. IV) which he "faute de mieux" finds in Reichenbach's "straight rule" (cit. loc., pp. 99, 114). The next logical step consists in "pragmatic retrojustification" (chap. V) which makes some circularity inevitable. Therefore he has to fight the "charge of question-begging" (chap. VII), his defense being that "what we have here is not a vicious circle, but an essential feedback mechanism" (p. 124). He summarizes his thoughts in a single illustration (Figure VII.1, p. 123) that indeed has the structure of a recursive algorithm:

(i) Prior justification of a certain method M

(ii) Application of M

(iii) Results R, obtained with the help of M

(iv) In particular, there are results on M's efficiency E

(v) Feedback: E provides an improved (posterior) justification of M

(vi) Continue with (ii)

Note, that this procedure is quite similar to the calculation of $n!$ in section 3.7. Yet, unlike an algorithm that stops after a finite number of steps, the procedure described by Rescher is open-ended: Running the program over again leads to improved justifications of M which could be summarized with the words "learning by doing". Naturally, a hierarchical structure thus develops (see Illustration 3.3, Skyrms (2000), chapter III.3, Braithwaite (1953), p. 263, and Kelly (1996), in particular pp. 64-68, 91, 116). The straightforward charge of infinite regress (Bonjour 1997) can be coped with easily. An infinite regress defers justification of some statement A endlessly:

$$\ldots \Rightarrow D \Rightarrow C \Rightarrow B \Rightarrow A$$

However, Rescher's procedure starts in the well-defined step (i), and may be applied over and over again. In particular, a certain method can thus be improved successively. In other words, there is room for improvement with the enhancements building on each other:

Prior justification $\rightarrow$ 1. improved justification $\rightarrow$ 2. improved justification $\rightarrow \ldots$

That's a benign and rational way to proceed. Justification is not postponed endlessly, rather, it builds up gradually. Thus, in this sense, M is self-supporting (Black's idea).

Following Achinstein (1961), philosophers seem to agree that Black did not succeed.[11] However, acknowledging that circularity need not be malign, it does not suffice to just point to circularity in order to dismiss some argument right away. By their very definition, iterative procedures refer back to themselves (are circular in this sense), yet it would be ridiculous to claim that any such procedure is useless. Benign recursive mechanisms, like those employed in the theory of recursive functions (cf. Kelly (1996), chapter 6.6), iterative calculations used throughout mathematics, or the procedure elaborated by Rescher, neither beg the question nor do they lead into an infinite regress. (Of course, in practice, convergent sequences need to be terminated at some point, but that neither impairs their practical value nor their theoretical merits.) While circularity involving some premise (e.g., $p \Rightarrow p$) is clearly begging the question, "rule induction" in the form presented by Rescher is logically sound.[12] Here is another simple argument in his favour: Successful ends justify the means by which they have been reached. Thus, if B inductively supports conclusion C, and C turns out to be true, this success also lends strength to the inductive leap by which C has been reached. Iterating this idea, many true conclusions of this kind support our trust in inductive methods in general.

Quite often, feedback loops are self-regulating. In the case of the above algorithm, it is straightforward to assume that successful inductive arguments are reinforced, i.e., that they will be used more widely, and that unsuccessful inductive moves are enfeebled and possibly abandoned. Thus induction also has an element of "self-correction" or "rational selection" to it (see Rescher (1980), in particular his chap. V, explicitly building on Peirce). In other words, an iterative procedure eliminating unsuccessful generalizations and highlighting successful inductive steps is a (virtuous) "feed-back cycle of self-conformation and self-improvement" (ibid., p. 126).

[11] Their exchange is reprinted in Swinburne (1974), chap. 8.

[12] Papineau (1992), p. 15, gives a general definition (emphasis in the original): "An argument is premise-circular if its *conclusion is contained among its premises*; an argument is rule-circular if it reaches the conclusion that a certain rule of inference is reliable by *using* that self-same rule of inference (see Braithwaite (1953), pp. 276-7; Van Cleve (1984), p. 558." While there is a consensus that the first kind of circularity is malign, it is difficult to reach a verdict with respect to the second kind of circularity.

Now, on a meta-level, if, typically, or at least very often, generalizations are successful, inductive thinking (looking for a rule to those many examples) will almost inevitably lead to the idea that there could be some (general) "inductive 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).[13]

5.2.2 Which kind of induction?

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 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.[14] The same holds with the improvement of a certain inductive method which is Rescher's example: 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 man, 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 step 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.[15]

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

[14] 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.

[15] 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 sets 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. Also see the end of the next section on this point.

"Inductive validation of induction" is tricky for another but quite related reason: A concrete inductive step generalizes a particular statement. Formally, an elementary inductive step is a mapping from some specific statement A_1 to some more general statement A_2. Thus, in a typical inductive line of argument, inductive steps of order 0 connect statements of ascending generality:

$$A_1 \overset{f_0}{\to} A_2 \overset{g_0}{\to} A_3 \overset{h_0}{\to} A_4$$

Iterating this procedure, the arguments of inductive steps of order 1 are inductive steps of order 0, i.e., these mappings operate on inductive rules of order 0:

$$f_0 \overset{f_1}{\to} g_0 \quad \text{or} \quad h_0 \overset{g_1}{\to} g_0 \ldots$$

Of course, this train of thought can be extended to inductive rules of any finite order k, thus building a hierarchy of inductive arguments: "... scientific inductive logic is seen not as a simple, homogeneous system but rather as a complex structure composed of an infinite number of strata of distinct sets of rules" (cf. Skyrms (2000), p. 37). If this edifice is to be logically sound, one must not build on the same level of abstraction or reverse the order of argument. That is, f_k must not operate on some g_k, on f_{k+1}, or on some function of f_k which seems to be the case in Braithwaite (1953), p. 277. Moreover, as we have seen, it is pointless to look for some limit f_∞ (the general inductive principle). Formally, that is quite obvious. However, it is very difficult - if not almost impossible - to describe such nuances verbally, which could be a major reason why this kind of reasoning never became popular.

In a sense, induction is like Munchhausen's swamp. If you only grab your own hair, you will never get out (viscous circularity). However, if you find, in addition to that, at least some foothold, you can start from there. Although each successive step may fail (due to the inevitable inductive gap), it is possible to slowly - step by step - improve your position (see the illustrations in section 3.3). Successive inductive statements rest on each other, and the charge of infinite regress can be coped with, in particular if there is just a finite number of iterations (see section 3.7), or in the case of "benign" convergence (section 3.5).

Black goes farther: Inductive investigation 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 A may borrow strength from funnel B (and vice versa), such that a conclusion resting on both lines of investigation may be further reaching or better supported. A classical example is Perrin (1990) on the various methods demonstrating the existence of atoms. In the allegory of the swamp, it suffices that one man of Munchhausen's company finds some hold, in order to help others out. With more and more men finding support, it gets successively easier to make progress. Figuratively, Illustration 3.3 is multiplied until the funnels begin to support each other on higher levels of abstraction:

[Illustration 4: Several funnels]

Combining all lines of inductive reasoning and their results, the complete edifice of knowledge indeed looks like a "tangled web" which, in a sense, is "self-supporting". To this end, it suffices that there is some starting point and that circularity is not perfect. Rather, the crucial point is that the gaps between premises and conclusions remain readily comprehensible, and several lines of inductive reasoning can be linked such that they support each other. In the allegory of the swamp, the men will hold on to those having a good foothold but not to those still deep in the mud, groups of men supporting each other will form, until, finally, the whole company has reached a more stable state.

That's all that can be done. However, if the men are reaching out to the ghost of the general inductive principle, or if they are longing for the moon, they go down, since "to place substance, ultimate existence, *outside the limits* of human cognition, is to leave us enough mental room to doubt anything" (Groarke (2009), p. 87, my emphasis).

5.3 Past, present, and future

5.3.1 A model from calculus

Here is another crisp mathematical model for discussing Hume's problem: Given the value of some function f at a point x_0, what can be said about $f(x_0+h)$, i.e., the value of the function "nearby"?[16] Since a function, in general, may be defined in a completely arbitrary manner, $f(x_0)$ and $f(x_0+h)$ can assume any number. That is, without any further assumptions, given $f(x_0)$, nothing can be said about $f(x_0+h)$, no matter how small h.

As a matter of principle, the values of the function f have to be connected somehow, there needs to be some link between them. An elegant way to get to $f(x_0+h)$ is to impose smoothness conditions on f. That is, the function must not fluctuate arbitrarily. Instead, the smoother f, the more can be said about $f(x_0+h)$. In the words of Knight (1921), p. 313:

> 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.

The concept of differentiability defines this idea more precisely: A function is said to be m times differentiable, if

$$f(x_0+h) = f(x_0) + f^{(1)}(x_0)h + \frac{f^{(2)}(x_0)}{2}h^2 + \ldots + \frac{f^{(m)}(x_0)}{m!}h^m + R_m(x_0+h)$$

where $f^{(k)}$ denotes the k-th derivative of f, $(1 \leq k \leq m)$. The remainder $R_m(x_0+h)$ not only disappears if $h \to 0$, but even $R_m(x_0+h)/h^m \to 0$ if $h \to 0$.

The smoother the function, i.e., the larger m, the better the approximation of f at the point (x_0+h), given the function (and its derivatives) at the point x_0. In particular, "rule induction" could start with the coarse approximation $f(x_0+h) \approx f(x_0)$, proceed to the linear guess $f(x_0)+f^{(1)}(x_0)h$, continue with the sophisticated three-term estimate $f(x_0) + f^{(1)}(x_0)h + \frac{f^{(2)}(x_0)}{2}h^2$, etc. If m becomes arbitrarily large, the inevitable error

[16] Or shortly afterwards, if x_0 is a point in time and $h > 0$

of approximation can be made infinitesimally small: Functions with derivatives of any order can be approximated to any degree of precision.

In other words, Hume's vague idea of some "uniformity of nature" can be made rigorous, and a precise mathematical model once again confirms our naive intuition: The smoother a situation, the better the prognosis of $f(x_0 + h)$, given the information in x_0. In the best case, this information suffices to predict some value "beyond" x_0, in the model $f(x_0 + h)$, without error. Typically, however, there remains some difference. If the function is "just" (one time) differentiable, the last equation becomes

$$f(x_0 + h) = f(x_0) + f^{(1)}(x_0)h + R_2(x_0 + h)$$

where $R_2(x_0+h)/h^2 \to 0$ if $h \to 0$. This means that in a neighbourhood of x_0, f can be approximated by a linear function (but not more precisely), since $f(x_0 + h) \approx a + bh$ with the constants $a = f(x_0)$ and $b = f^{(1)}(x_0)$. If f is not differentiable at x_0, the function may still be continuous, i.e., $\lim_{h\to 0} f(x_0+h) = f(x_0)$, meaning that $f(x_0+h)$ is not completely arbitrary if h is small: For any sequence $(x_i)_{i\in\mathbb{N}} \to x_0$ one also has $f(x_i) \to f(x_0)$. (However, the rate of convergence can be arbitrarily slow.) Without continuity, i.e., any substantial assumption, information about f at x_0 does not give any clue about $f(x_0+h)$. So, in this model, it's the weak assumption of continuity that makes predictions possible.

5.3.2 Will: The moving boundary

Interpreting Hume's problem in an empirical way is straightforward. Hume (1740/2010) writes about the uniformity of nature as follows (emphasis in the original):

> All probable arguments are built on the supposition that there is this conformity betwixt the future and the past, and therefore can never prove it. This conformity is a *matter of fact*, and, if it must be proved, will admit of no proof but from experience. But our experience in the past can be a proof of nothing for the future, but upon a supposition that there is a resemblance betwixt them.

That is also the starting point of Will (1953). However, he first discusses a useful spatial analogy (p. 41-42):

> Suppose that there was somewhere in the world an enclosure beyond which it was impossible for anyone ever to go or to make any observations. Nothing could be seen, heard, or in any other way perceived beyond the border.

Obviously, this is quite a reasonable model for the land of the dead. Without any kind of information flowing from there to the land of the living (e.g., some kind of resurrection), nobody will ever learn anything about Hades. In Hume's problem, however,

> The territory beyond the enclosure, for ever barred from human perception, is the land of the Future. The land within the enclosure is the land of Present and Past...

In other words, the standard model considered is static: On the one hand there is the Past up to the Present, on the other hand the land of Future, both separated by an insurmountable barrier. Will readily notices that this is an inadequate model (p. 44):

> If the border had not yet begun to recede [the inhabitants of Past] would indeed be in an unfortunate position... But this is not the case. The border is constantly receding. And granting that it will constantly recede, revealing always more of the land of Future, and even granting also that this means that there is an inexhaustible area to be revealed, the inhabitants of Past are in the fortunate position ... that they may learn more and more about chickens [or any other phenomenon], Past and Future.

The crucial point is that the empirical situation is not static but dynamic, and thus much more fortunate for some observer than the sceptics' scenario. The receding boundary constantly generates information that enables us to go well beyond the Past. Arguing within the framework of the last section, it's the dynamics that create a link between $f(x_0)$ and $f(x_0+h)$. In the discrete model, we may not only learn something about nature every single day, but also about the transition from day i to day $i+1$. In general, static and dynamic properties of the world may be studied in a rational manner.

A standard retort of the sceptic could be that this is all well and good, as long as there is no qualitative change. Permanent sampling might give us some clue, but what if the future turned out to be completely different from the past? In Russell's words:

> We have experience of past futures, but not of future futures. Will future futures resemble past futures? ... We have therefore still to seek for some principle which shall enable us to know that the future will follow the same laws as the past. (Russell (1912), pp. 100-101, cited in Will (1953), p. 45)

No. Here is a concrete example: Given a stable large-scale weather pattern, we may collect data and predict tomorrow's weather precisely. Russell says that we must fail if there is a (sudden) break in the weather. But this is not true. Although a qualitatively different pattern complicates prediction (in the worst case devaluing all observations so far), the overall situation is still fortunate for the observer: Future becomes Past, generating information. For example, given a new large-scale weather pattern, the observer may proceed as before and learn more and more about this situation. Of course, estimating a volatile situation (e.g., several patterns with sudden swings between them) is more difficult than approximating a stable setting. However, information accrues nevertheless, and it is thus rational not only to estimate tomorrow's weather but also the next large-scale weather pattern and when the swing will occur.[17] In a nutshell, "uniformity of nature" is not necessary for rational estimation. Even if there is a sudden break (change of law), rendering all information gathered so far useless, the arrow of time constantly provides the observer with new evidence.[18]

The example of weather prediction demonstrates that a finite set of large-scale patterns (attractors, lawlike situations) with sudden jumps between them does not preclude rational prediction. The only scenario removing the future effectively from the observers'

[17] In statistics one therefore distinguishes between several classes of parameters, nested and hierarchical models etc. For example, it is straightforward to estimate the parameters a_i of an ordinary regression, i.e., to minimize the expression $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i(x_i))^2$ where (x_i, y_i) are the observed values $(i = 1, \ldots, n)$, and $\hat{y}(x) = a_0 + a_1 x + a_2 x^2 + \ldots a_k x^k$. It is more difficult, however, to estimate k, the number of parameters (the complexity of an appropriate model) also.

[18] However, a very real problem, not considered by Hume, is that information may decay. We know much about the present, less about the recent past, and hardly anything about bygone days. Destroying a historic source means that it is gone forever.

eyes are *permanent qualitative* changes. World history of the last few centuries provides an example: Because of the considerable stream of scientific breakthroughs, technological revolutions, social change and political turmoil, it would have been almost impossible for an educated person living several hundred years ago to predict what the world looks like today. Used to carriages, he may have thought of improved coaches, automobiles, perhaps. However, nobody had a clear idea of aircraft, ballistic missiles, and space travel. For another illustration, see Black (1954), chap. 12.

In a nutshell, the dynamic scenario, constantly rendering Future into Past, provides an attentive observer with an endless stream of information. This model is appropriate for the world we live in, and it strongly favours rational predictions. They are the rule, not the exception. Thus, again, critics need to fall back on extremely volatile situations. However, there is a large difference between "conformity" in the sense of "there are no extremely pathological fluctuations" and "conformity" in the sense of stationarity or far-reaching resemblance, which could be closer to Hume's original idea.

5.3.3 A useful distinction

A corollary to the dynamic model just discussed is that any definite point in the future sooner or later becomes past, and thus accessible. The only point that remains forever hidden is the point ∞, by definition exceeding any finite boundary. Therefore Will (1953) makes a useful distinction: When speaking of 'future instances' he advises us to distinguish between 'future-1', which qualifies specific events and things that are currently future but will eventually become past, and 'future-2', which describes in the abstract that portion of the space-time universe which is always "beyond the line of the moving present".

He goes on to argue that Hume must have used 'future-2' when he says that we cannot know that the future will resemble the past, for we are in fact continually confirming that 'future-1' instances resemble past instances (i.e., whenever a 'future-1' instance finally becomes a present observed instance). Yet 'future-2' instances are by definition unobservable — and we ought not to be in the least concerned about that trivial truth.[19]

In essence, Will thus distinguishes between the well-defined empirical problem and a badly defined formal problem. In the real world, we continuously gather information. Thus the more we have learned about a certain situation, the better we are able to predict what is going to happen. Also appropriate formal models confirm (in particular see sections 3.6 and 5.3.1), that there is nothing mysterious about this, as long as the whole framework is bounded.

Future-2, however, introduces an infinitely large gap, since even in principle, the day when pigs fly is beyond our reach. Including future-2 can only be done in a formal situation, and in doing so one has to take pains to guarantee a bounded gap. As we have stated before, carelessly dealing with an idealized formal model and using highly

[19]Referring to Russell (1912), Edwards (1974), p. 38 (emphasis in the original), calls this move "an *ignoratio elenchi* by redefinition, [since] in its ordinary sense, 'future' does *not* mean 'period which has to the past and the present the relation of happening after it *and* which can never itself be experienced *as a present.* [However,] if the word 'future' is used in its ordinary sense in the statement 'the future will resemble the past and the present in certain respects', then we have plenty of evidence to support it."

empirical language/content is dangerous, for it may result in a confusing situation, in an opaque problem, having no reasonable answer:

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

In this sense, Hume's paradox arises since we confuse a well-defined, restricted situation (line 1 of the table above) 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 or some other model described above), 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.[20]

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 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 (line 1 in the above table), but not accept any framework (line 2 in the above table). Arguing "off limits" (more precisely, beyond any limit restricting the situation somehow) can only lead to a principled and completely negative answer.

It is rather ironic that a relentless critic of induction, not trusting induction at all, thus reaches the same conclusion as an adamant inductivist (being led to some general, yet non-existent inductive principle, see section 5.2.2). Although their starting points as well as trains of thought are completely different, they arrive at the very same place - the void.

5.4 Williams: Sample and population

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

[20] 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."

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.

5.4.1 Random sampling

Williams's idea is striking. The most straightforward criticism focusses on the link between population and sample. For ease of discussion, suppose the population P is an urn containing red and white balls, the proportion of red balls in the urn being $r\%$. Critics then argue among the following lines:

(i) Implicitly, Williams supposes that the next ball to be drawn is red with the probability $r\%$. This is an implicit symmetry assumption that need not be justified. More generally speaking: Couldn't there be a demon favouring some of the balls?

(ii) Doesn't "selection at random" presuppose some notion of randomness? In other words, is there some kind of circularity involved?

These critics are right to complain that the necessary link between sample and population P is rather implicit. Williams may not have made it explicit, since the symmetric link between population and sample is standard in probability theory. Moreover, it should be noticed that we are discussing a theoretical model here, and within this model, the assumption is very natural.[21]

Of course, if $0 < r < 1$, the ways in which the demon D could alter the selection of the balls are almost endless. In principle, if his way to change the setup of the sample is traceable, the effect is as if the balls of the sample were drawn from a different population $P^* = D(P)$. For example, if the demon constantly prefers red balls, the observable ratio $r^*(S_n)$ will be augmented and quickly converge towards $r(P^*)$. If the demon acts in a chaotic way, however, it is no longer possible to track the consequences of this mechanism analytically. For example, it suffices that the demon's colour preference changes periodically (with the periods becoming longer at a certain rate) in order to have $r(S_n)$ fluctuate but not converge. For these reasons, theory decides to tackle the simple (and because of the foregoing explanation rather general) situation of a random fair selection of balls.

"Random" here means: Each ball has the same chance of being selected, and there are no dependencies among subsequent balls being chosen. This being said, the "hyperpopulation" (Williams's terminology) of all samples S_n of size n is just what statisticians call the "sample space", and the larger part of Frequentist statistics is based on this space of all possible samples. In the simplest cases, like that of Williams (1947), p. 97, these inferences boil down to combinatorics, i.e., a counting argument. So the notion of "randomness" is by no means circular, rather, it is a well-defined logical probability.

[21] As a matter of fact, it is so much straightforward that it is not even mentioned.

5.4.2 Asymptotic statistics

Though much discussion has been going on in philosophy (for a review see Stove (1986), Campbell (2001), Campbell and Franklin (2004)), it is quite obvious to a mathematician that Williams's theoretical reasoning is sound:

Informally speaking, the gap between sample and population isn't merely bounded, rather, it is quite narrow from the beginning. Moreover, if the size of the sample is increased, it can 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. In general, of course, the information in the sample almost equals that in the population only if certain conditions hold, i.e., if the link connecting population and sample is not pathological. For example, the combinatorial argument (most large samples are typical) only holds if the selection of these samples is unbiased. If there is, like in Indurkhya (1990), an unknown or untraceable mechanism D preferring some samples (e.g., extreme ones) over others, we cannot infer anything about the true population P, but all about $D(P)$. In a nutshell, it takes a really witty and malicious demon to prevent convergence towards a reasonable number.

Many philosophers have focused on details of Williams's example. However, that is quite unnecessary, since the formal support in his favour is overwhelming: The very core of asymptotic mathematical statistics and information theory consists in the comparison of (large) samples and populations. Their most basic theorems say that, given mild conditions, samples approximate their populations, the larger n the better: The main theorem of statistics guarantees convergence of the empirical distribution function toward the population's distribution function. Several limit theorems, in particular the Central Limit Theorem, state convergence in a probabilistic sense. Similarly, in information theory, the asymptotic equal partition property (see section 3.6) is a main result to the positive.

Williams's very specific example works due to laws of large numbers which guarantee that (most) sample estimators are consistent, i.e., they converge toward their population parameters. In particular, the relative frequency of red balls in the samples considered by Williams converges towards the proportion of red balls in the urn. That's trivial for a finite population and selection without replacement, however convergence is guaranteed (much) more generally. Convergence is also very robust, i.e., it still holds if the seemingly crucial assumption of independence is violated (in fact, non-existent, see Fazekas and Klesov (2000)). Moreover, the rate of convergence is very fast (see the last verbatim quote, and the literature building on Baum et al. (1962)). Extending the orthodox "two component" (sample-population) model to a Bayesian framework (consisting of the three components 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.

But also pathological examples are quite interesting. When is it difficult to proceed from sample to population, or more crudely put, from n to $n+1$? 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 viscous

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, almost nothing can be said about θ since $\sum_{i=1}^{n} x_i$ may still be close to zero.

Vicious 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 (e.g that there is no rational basis for any inductive claim). It needs an intrinsically difficult situation and an "unfair" selection procedure to obtain a systematic bias rather than to just slow down convergence. In a nutshell, statistics and information theory both teach that - typically - induction is rationally justified.

It should be mentioned that the concepts of information and probability can often be used interchangeably. However, a major technical advantage of information over probability and other related concepts is that the information of any observation in non-negative. Thus information monotonically increases.[22] In the classical "nice" cases, information accrues steadily with every observation. So, given a large enough sample k, much can be said about the population, since the difference $I(n) - I(k)$ is rather small. More 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 on 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.

5.4.3 Sampling in practice

Giaquinto (1987), p. 614, rightly admits that "a lot depends on how the sample is obtained. If we know that our sampling procedure is truly random so that we are just as likely to end up with one (large) sample as any other sample of the same size, we may reasonably infer that the sample obtained is probably representative." This being said, and this being guaranteed in the theory of random sampling, he explicitly, and all of a sudden, turns to real-life samples: "Suppose now that a high proportion of ravens are not black but all of these live in very remote regions, are difficult to spot, and are easily mistaken for birds of another species when they are spotted. In this circumstance we are likely to obtain a sample which is biased in favour of black; our sample is probably unrepresentative, even though most large samples are representative."

As a matter of fact, he is perfectly right: A "convenience sample", obtained by some obscure or completely unknown procedure, can be thoroughly misleading. Brown (1987), pp. 117-118, also remains on track when he says that, *in practice*, "we are given absolutely no grounds for the assertion that our observations are gathered through a fair sampling procedure." However, that is exactly why, given the above theory and being

[22] This is not the case with probabilities: Adding a relevant condition may increase or decrease the value of a conditional probability. However, Williams's convergence argument "sample → population", and the asymptotic theorems of probability theory (laws of large numbers, etc.), of course, still hold. As stated in section 3.5, non-monotonicity does not render a convergence argument invalid.

well-aware of biased samples, statisticians invented experimental design. That is, they take explicit measures to get as close as possible to the ideal of random sampling. Following Fisher (1935/1966), random procedures controlled by the experimenter have become the gold standard in applied statistics, distinguishing (real) "statistical experiments" and "quasi-experimentation" (Shadish et al. 2002). In other words, since statisticians know very well that representative samples are not guaranteed, they work hard to obtain them.

It may be added that, nowadays, much data is collected automatically ("big data"). Although not quite random samples, these stacks of data contain a lot of information, and it would be irresponsible to ignore this source of knowledge. Although opposed by orthodox statistics (being committed to Fisher's heritage), "exploratory data analysis", "data mining" and their ilk have become popular and successful. Theory gives a reason why: Since convergence is fast, most samples are, if not representative, still quite typical for some population.[23] It takes a strong force, e.g., a heavily biased selection procedure, to get a distorted picture of reality. However, the stronger such a nuisance influence is, the more obviously it leaves its traces in the data and elsewhere, and the more easily it can be detected, and possibly corrected.[24]

Be this as it may, the crucial point in discussing random selection is that one must not confound theory and practice: Theory studies interesting natural models, random selection being among statistics' most popular mechanisms. Given such a strict formal framework, it can be proved that most (large) samples are similar to their population. So, theoretically, there are no grounds for fundamental doubt. In empirical work however, we typically do not know if nature provided us with a representative sample. Thus, in real-life, there are often grounds to question the assumption of random sampling mechanisms. Since it is indeed dangerous to generalize from a convenience sample, appropriate measures have to be taken, and the decisive ingredient of any traditional statistical experiment is a random mechanism controlled by the experimenter. This application of the theory of random sampling then provides the ground for external validity, i.e., generalizability, of the trial.

5.4.4 Unknown sampling mechanism

The last line of attack remaining is thus purely empirical: In most significant real-life situations where we use inductive reasoning, we do not know whether sampling was fair or not, so most of our important inductive inferences are not justified. What can be said about this argument?

First, theory shows that the odds are highly stacked in favour of a typical sample. Moreover, it teaches that typicality is a robust property, i.e., it takes a lot to make a sample completely misleading. Thus, unless there is some specific reason to suspect that a given sample is non-representative, one should assume that it is. The burden of proof is with the critic, not with the researcher trusting his sample by default (Campbell and Franklin 2004). As a rule, one should expect that most empirical samples are somewhat distorted but still give valuable information about the situation from whence they they come. Typically, we can learn something about the world when we listen to

[23] Epidemiologists could mention the Framingham study at this point, see Greenland (1990).

[24] It may be added that even if large convenience samples are systematically biased or unreliable in other respects, it pays to appreciate what they have to say (with a reasonable amount of doubt), since a contaminated source of information may be better than none.

the data. If this were different, fields like astronomy, archeology, or epidemiology, all of them mainly based on somehow received and at times highly selective data, would be virtually impossible.

Second, conclusions reached with the help of one sample may readily be checked with another sample. Replication in particular and cross-validation in general are highly recommendable scientific procedures. It can be shown in theory (e.g., in the derivation of the normal distribution) and observed in practice (e.g., Perrin (1990)) that errors may cancel each other out, leaving behind a clear signal. In general, borrowing strength is typical for inductive inferences (see Illustration 5.6 toward the end of section 5.2.2).

Third, if a demon contaminates a certain sampling procedure, he leaves traces in the data. Thus, statistical methods and cross-validation may also be used to uncover and pin down the demon (for example, to isolate an important nuisance factor). In short, an empirical sampling mechanism can be investigated like any other natural phenomenon. Since there are methods to uncover how it works, empirical demons cannot hide forever.

Fourth, although each bit of data can be interpreted as a kind of sample from some population, this model should not be overused. Of course, it is important to know where data come from, how they were collected or compiled. However, this model is also rather narrow. For other kinds of situations, different classes of models are more useful and in fact available. For a general treatment of the formal and the empirical inductive problem(s) see the sections above.

5.4.5 Maher's criticism

Maher (1996) chooses yet another strategy. In addition to the population and the sample, he invokes a prior distribution. Within this extended setup, his argument runs as follows: Suppose the population consists of birds, most of which are black. Does a large sample with 95% black birds sustain the claim that the population is mainly black? Williams argued for a clear "yes". Maher, however, says that "no other relevant information besides the sample" should be formalized by a prior distribution on the set C of all possible colours. Since there could be birds of any colour, "no knowledge" may be represented by a rather large set C, with the prior rate of the colour black being very small. Of course, given such a prior and a typical sample, the posterior rate of the colour black increases but may still be considerably smaller than the proportion observed in the sample. Thus we are not entitled to conclude that most birds of the population are black.[25]

Why is this example correct yet nevertheless pathological? Straightforwardly, given just the sample-population model, the sample is very close to the population, and Williams's conclusion is rationally justified. In order to undermine this idea, Maher has to make the inductive gap large. Thus he introduces a prior [P(black) close to zero] that is far away from the population [due to the sample, P(black) is very likely

[25] In his example there are $k = 145$ different colours i, black being one of them. If nothing else is known, by the principle of indifference, $P(i) = 1/145 \approx 0.007$ for $i = 1, \ldots, 145$. The data consists of $n = 3020$ birds, b=2869 of these being black (Stove 1986), thus $b/n \approx 95.0\%$. In this case, the Bayesian mechanism shows up as conjugate distributions telling us that the posterior distribution of the colour black is a certain Beta distribution, i.e., $\beta(1+b, k+n-(1+b)) = \beta(2870, 295)$, giving the predictive proportion $P(\text{black}|\text{prior, sample}) = (1+b)/(k+n) = 2870/3165 \approx 90.7\%$.

close to one].[26] The sample, although large and most probably representative, cannot change the proportion that much, i.e., the estimated rate of the colour black, being a convex combination of the prior proportion (0.7%) and the portion of black birds in the sample (95.0%) remains rather low:

$$\hat{P} = 90.7\% = \lambda \cdot 0.7\% + (1 - \lambda) \cdot 95.0\%$$

Consistently, the inductive step from there to the population remains considerable, and he summarizes that we should have doubts in the conclusion that the population mainly consists of black birds - although 95% of all birds observed were black!

However, though Maher merely claims to model "ignorance" within a Bayesian framework, and though, mathematically, the weight of the prior in his numerical example is just $\lambda = k/(k+n) \approx 0.046$, the example gives too much weight to the prior and too little weight to the sample. That is, his final observation (a large difference between his estimate and the true proportion) crucially depends on the prior. This emphasis would be justified if we knew a lot about the population, i.e., if we were already "close" to the truth, and the purpose of the sample were to "fine tune" our knowledge (see the last line of the next table). However, since the prior gives a lot of colours equally small weights, it represents just the opposite - ignorance. Given this starting point, far away from the probable truth of a homogenous population (most birds having the same colour), any typical sample will move the prior belief in the right direction, but not enough so, resulting in a prior-data conflict.

That's why statisticians prefer "to let the data speak for themselves" if no substantial prior information is available. That is, they give very little weight to the prior, and much weight to the empirical evidence available in the data. (In other words, their conclusions mostly depend on the data.) Frequentist statisticians straight away work without a prior, Bayesian statisticians wait until the prior has "washed out", i.e., until enough data has been collected to "overwhelm" the non-informative prior.[27] The following table may illustrate this, and further demonstrate that Maher's numbers are artificially chosen:

Situation	k	n	b	λ	Estimated proportion $\hat{P}$	Difference $95\% - \hat{P}$
Sample only	-	3020	2869	0	95.0%	0
Maher's example	145	3020	2869	0.046	90.7%	4.3
Fewer observations	145	302	287	0.324	64.4%	30.1
More observations	145	10000	9500	0.014	93.7%	1.3
Meager prior information	1450	3020	2869	0.324	64.2%	30.3
Considerable prior info.	15	3020	2869	0.005	94.6%	0.4

In a nutshell, Maher's example confounds two completely different situations: Either your (prior) knowledge is rich, then you should give it a lot of weight, or it is poor, then it is advisable not to rely too much on it. However, it is neither reasonable to ignore substantial prior information (a standard criticism of Bayesians toward Frequentists), nor to count on feeble prior information (the Frequentists' retort). Maher goes even farther and bases his final conclusion on *non-existent* prior information. Campbell

[26] According to Maher's correct calculation, given just the sample, the probability that the true proportion in the population lies in the interval $95\% \pm 3\%$ is 0.999.

[27] See the first and the fourth line of the next table.

(2001) calls this absurd and gives a nice example to illustrate matters (emphasis in the original):

> ... suppose there is a mass of evidence that points overwhelmingly to the guilt of Brown for the murder of Smith. *A priori*, it is very unlikely that Brown killed Smith, because there are so many people who could have done it. But we do not say, 'Despite the overwhelming evidence against Brown, we should not suppose that he killed Smith, because it is *a priori* unlikely that he did'. No-one would take such reasoning seriously.

Unfortunately, Maher's overall strategy is quite typical for a number of authors. Instead of adhering to a certain model, within which induction is rational, they *extend* the situation and try to show that induction does not work there. In principle, this is a rather unfair strategy, since it evades the challenge of responding to the situation at hand. In mathematics, counterexamples are only accepted if they address a well-defined *given* situation. One is not allowed to switch horses (discuss a different, in particular a more general situation) for the obvious reason that a theorem may not hold outside the domain in which it was proved.[28] Philosophers rather work with verbal arguments, and it is easier to switch horses in that realm. For example, there is no escaping the fact that Williams is formally right. But instead of acknowledging that a convincing priori justification for induction has been found, many authors readily move on to empirical samples, or discuss extensions/related situations, quite a few of them rather pathological. Following this strategy, rather weak counterexamples evolve, the broad range of mathematical results in favour of Williams is ignored, and the discussion wanders off course.

5.5 Statistics - old and new

The computer sciences (in particular, see sections 3.1, 3.2, and 3.8) and statistics have particularly elegant models to deal with the problem of induction. Their key ingredient is information:

> The object of statistics is information. The objective of statistics is the understanding of information contained in data. (I. and M. Miller (1994) in Barnett (1999), p. 3)

Observations add information. The most straightforward way to model this is an interval $[a, b]$, having a certain width $b - a$, that becomes smaller when data is added. A lower and an upper *prevision* (cf. Miranda and de Cooman (2014)), defining $[p_l, p_u]$ are a contemporary and particularly elegant, but not yet widely used, way to handle such matters. Within this framework, one starts with a "prior near-ignorance model" which in the example of the last section would be the interval $[0, 1]$. The posterior interval for p, the proportion of black birds in the population, is defined by $p_l = b/(n+m)$ and $p_u = (b+m)/(n+m)$, where $m \in \mathbb{N}$ defines the weight of the prior.[29] The standard choice $m = 2$ yields $[94.9\%, 95.0\%]$, $m = 100$ gives $[92.0\%, 95.2\%]$, and $m = 1000$ leads

[28] For example, a counterexample to Pythagoras' theorem would have to start with a right-angled triangle, since the theorem does not hold for general triangles.

[29] That is, m can be interpreted as the size m of a (hypothetical) sample that was obtained before the current sample of size n at hand.

to $[71.4\%, 96.2\%]$. A nice feature of this approach is its independence of the number of colours k considered (cf. Augustin et al. (2014), pp. 151, 159-161).

5.5.1 Adding information

To date, the formula arguably most often used is

"Prior information + Information in some set of data = Posterior information"

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

$$\begin{aligned} I(H,\mathbf{x}) = I(H) + I(\mathbf{x}|H) \quad &\Leftrightarrow \quad -\log p(H,\mathbf{x}) = -\log p(H) - \log p(\mathbf{x}|H) \\ &\Leftrightarrow \quad p(H,\mathbf{x}) = p(H)\cdot p(\mathbf{x}|H) \end{aligned} \tag{3}$$

In other words: The first level consists of the prior hypothesis H, the second level 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 coarse picture becomes more detailed. Notice, that (3) 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 can be just *one* 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 (3). Traditionally, the gambling schemes just introduced are called 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 which is also reflected in (3).

5.5.2 Both tiers well-defined

In Bayesian statistics, the upper and the lower level (containing more / less information) are both explicitly modelled by a parameter θ having some distribution D_θ. Typically, the level of knowledge is measured by the variance or the entropy of D_θ with the data reducing this variance. The upshot is that the "data update" connecting prior and posterior distribution amounts to an application of Bayes theorem and can thus be done in a straightforward way. In this narrow sense, Bayesians solve Hume's problem (Howson und Urbach 2006). If prior and posterior are conjugate distributions, i.e., if they belong to the same family of distributions, "Bayesian updating" is particularly elegant. For example, suppose $\mu \in [a,b]$, and every value of μ is equally likely.[30] If

[30] That is, $\mu \sim U[a,b]$ where $U[a,b]$ is the uniform distribution on the interval $[a,b]$.

each observation is modelled by a normal distribution $X_i \sim N(\mu, \sigma)$ with the standard deviation σ known, n independent observations lead to the posterior distribution $\mu \sim N(\mu, \sigma/\sqrt{n})$. Not surprisingly, however, problems occur if the first layer, i.e., in this model, the prior distribution, is not a proper probability distribution. "Improper priors", i.e., measures with an *infinite* mass, have plagued the field for decades.

A more recent application of (3) is the decomposition of data into pattern + noise. In other words, the inductive task is to find the (relevant, non-transient) information hiding in the data. Simple hypotheses do not contain much information, and nor do the data, if they are well approximated by some hypothesis. In other words, it is reasonable to minimize (3), the optimum being attained by a rather simple hypothesis fitting the data well. This criterion is called "minimum message length (MML)" (Wallace 2005). If $I(\cdot)$ is replaced by Kolmogorov complexity $K(\cdot)$, one obtains "minimum description length (MDL)" (Rissanen 2007). Since $K(\mathbf{x})$ focuses on the complexity of (information contained in) a *single* string of data, MDL is very general and adapted to the data at hand.[31]

5.5.3 Upper tier well-defined

In Bayesian statistics the prior distribution seems quite arbitrary, in particular if there is no specific prior information. Therefore Frequentist statistics has tried to avoid this commitment. Instead, Fisher argued as follows: In the example just described the true value $\mu_0 \in \mathbb{R}$ is unknown (i.e., μ_0 is a fixed but otherwise *arbitrary* real number). All possible (independent) samples of size n constitute the sample space and if $X_i \sim N(\mu_0, \sigma)$ with σ known, we have $\bar{X} \sim N(\mu_0, \sigma/\sqrt{n})$ where $\bar{X} = (X_1 + \ldots + X_n)/n$. What can we thus say about μ_0?

Since, given μ_0, the distribution of $\bar{X}$ is known, we can calculate the probability of a small deviation $P(|\bar{X} - \mu_0| \leq \epsilon)$, i.e., that the value of $\bar{X}$ departs from μ_0 by no more than ϵ. In particular, we may choose a number $k > 0$ such that $P(|\bar{X} - \mu_0| \leq k\sigma/\sqrt{n}) \geq 0.95$. In other words, choosing k in this way, $\bar{X}$ and μ_0 will be "close" in at least 95% of all cases. Given the normal distribution, it turns out that $k \approx 2$.

In other words: In the hyper-population of all samples (to use Williams's terminology), the proportion of good samples is at least 95%. Since

$$\begin{aligned} P(|\bar{X} - \mu_0| \leq 2\sigma/\sqrt{n}) \geq 0.95 \quad &\Leftrightarrow \quad P(\bar{X} \in [\mu_0 - 2\sigma/\sqrt{n}, \mu_0 + 2\sigma/\sqrt{n}]) \geq 0.95 \\ &\Leftrightarrow \quad P(\mu_0 \in [\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n}]) \geq 0.95, \quad (4) \end{aligned}$$

Fisher said that the odds are at least 19:1 that the arithmetic mean $\bar{x} = (x_1 + \ldots + x_n)/n$ of a concrete sample $x_1, \ldots, x_n$ is close to μ_0. The move from the sample space to a concrete sample - i.e., sampling - thus defines a logical probability. In the words of Hampel (2005): "[A] scientist using statistics can bet 19 : 1 that the unknown fixed parameter is in the fixed (but randomly derived) confidence interval..." Notice, however, that for exactly the same reason we do not know if we got a typical (good), or an atypical (bad) sample. All we know is that with "sampling probability" of at least 95%, the true value μ_0 lies within the concrete confidence interval, and with at most 5%, the true value is located *somewhere outside* this interval.

[31] For many details see Li and Vitányi (2008), in particular their chapter 5.4., where these authors connect this kind of reasoning with the logarithmic version of Bayes' rule.

Neyman, however, argued as follows: Given a concrete sample, the fixed number μ_0 either lies in the interval $[\bar{x}-k\sigma/\sqrt{n}, \bar{x}+k\sigma/\sqrt{n}]$ or not. That is, $P(\mu_0 \in [\bar{x}-k\sigma/\sqrt{n}, \bar{x}+k\sigma/\sqrt{n}]) = 0$ or $= 1$. In other words, given $\bar{x}$, a specific realization of $\bar{X}$, there is no probability. In this view, restricted to the sample space and functions defined on that space, only the random variable $\bar{X}$ has a probability distribution, neither does the realization $\bar{x}$ nor the parameter μ. Thus neglecting Hampel's remark "(but randomly derived)" in brackets, leads to a cautious and rather "objective" standpoint. However, it is also a "view from nowhere", since nobody knows whether the probability in question is 0 or 1. An investigator's level of information rather corresponds to Fisher's who (at least in this case) gives a sound reason why the researcher's intuition is correct: Even after seeing the data, a certain "subjective" amount of uncertainty remains which can be quantified with the help of the above probability statement.

The main reason why Fisher's idea did not succeed is the fact that the last equivalence in (4) is not trivial. In general, one *cannot* swap probability from some statistic to some parameter (from $\bar{X}$ to μ in the present example) which is the crux of his "fiducial argument", the consensus now being that this idea was Fisher's biggest blunder (Efron 1998).[32]

The basic problem with Fisher's general approach is that the tier containing less information is somewhat vaguely defined (μ_0 being *some* real number is very close to an implicit improper prior). Despite this "loose end" one is able to derive reasonable results in well-behaved cases, in particular, if a single parameter of the normal distribution has to be estimated. However, the idea breaks down almost immediately, if the situation is a bit more complicated, e.g., if both parameters of a normal distribution are unknown (the so-called "Behrens-Fisher problem", see Pitman (1957)), or if the normal is replaced by some "wilder" distribution (e.g., the Cauchy, see Jaynes (2003), pp. 502-503).

5.5.4 Connecting finite and infinite sequences

In a sense, statistics' frameworks are rather complicated. A more straightforward formal model consists in 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. Thus we have two well defined layers, and we are in the mathematical world.

Without loss of generality, Li and Vitányi (2008) study finite and infinite binary strings, i.e., $x_i = 0$ or 1 for all i. Their basic concept is Kolmogorov complexity (see section 3.8), not least since with this ingenious definition, the link between the finite and the infinite, becomes particularly elegant: First, some $\mathbf{x} = x_1, x_2, \ldots$ 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 (ibid., p. 230). Second, it could be proved that $\mathbf{x}$ is random if and only if there is a

[32] Since the interpretation of a confidence interval is somewhat tricky, many authors resort to the formulation that given many samples of size n, we can expect at least 95% of them to contain the true parameter μ_0. For example, Efron (1978), p. 234, writes: "...the interval $[\bar{x}-2\sigma/\sqrt{n}, \bar{x}+2\sigma/\sqrt{n}]$ covers the true value of μ with frequency 95% in a long series of independent repetitions of $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$". (Note that this statement somewhat confuses the concrete sample $\bar{x}$ and the population level, i.e., the random variable $\bar{X}$ having the distribution $N(\mu, \sigma/\sqrt{n})$.)

constant c, such that for all n, $K(\mathbf{x}_n) \geq n - c$. That is, "random sequences are those sequences for which the complexity of each initial segment is at least its length."[33]

Kelly (1996) devoted a whole book to the model described in the first paragraph. He also uses the theory of computability (recursive functions), but 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 quite surprisingly, he needs a strong regularity assumption, called "completeness", to get substantial results of a positive nature. In his own words (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."

Demanding less, i.e., 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), induction becomes much easier to handle. In this setting, it turns out that countable additivity of probability measures (see section 3.5) is a crucial regularity (continuity) condition, since then 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 (Walker 2003, 2004).

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 $\mathbf{x}$ satisfies certain *mild* conditions, the relative frequency f_n of the ones in the sample $\mathbf{x}_n$ *converges* (rapidly) towards f, the proportion of ones in the population. Our overall setup explains why:

First of all, there is a well-defined, constant population, i.e., a simple upper tier (e.g., an urn with a certain proportion p of red balls). Second, there is random sampling (see section 5.4.1), connecting the tiers in a consistent way. In particular, the probability that a ball drawn at random has the colour in question is p (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 $p\%$ red balls.

[33] Ibid., p. 221. The basic idea is to identify incompressibility with randomness, since any regularity is a kind or redundancy that can be used to compress a given string. The last theorem states that random (infinite) sequences can be characterized as being "completely incompressible", in the sense that any initial segment $\mathbf{x}_n$ is not compressible. The complete theory is developed 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.

(More precisely: The probability of all sequences containing about $p\%$ 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 $n = 1$.

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. Quite the opposite: The asymptotic theory of random processes is very beautiful and mathematicians have detected many limit theorems, even if sampling is not perfectly homogeneous. In particular, there are laws of larger numbers, guaranteeing the convergence of relative frequencies.[34]

5.5.5 Reichenbach's pragmatic justification of induction

As far as I see it, the pragmatic justification of induction tries to give a constructive answer based on very weak assumptions. Starting with a finite sample $\mathbf{x}_n = (x_1, \ldots, x_n)$, it is of course possible to calculate the proportion f_n of the symbol 1 in this vector, i.e., $f_n = \sum_{i=1}^{n} x_i/n$ is the observed relative frequency of ones. Since, empirically, most relative frequencies stabilize when n gets larger, Reichenbach (1949) considers some "practical limit" of f_n with n large enough.

Alas, this hardly leads anywhere, since f_n need not converge (e.g., the sequence f_n may finally fluctuate between two numbers). Even if $\lim_{n\to\infty} f_n = f$ exists,[35] a sequential observer, knowing arbitrarily long vectors $\mathbf{x}_n$, cannot decide whether the sequence has reached its limit, since the sequence may stay close to a "pseudo-limit" for any - but still finite - amount of time. Moreover, convergence can be slow, i.e., no matter how large n, f_n may still be far away from its true limit. At least, if f exists, one has $f_n \to f$ (rather by definition than by some law of large numbers), and thus, in this sense, it suffices to consider frequencies, instead of other, possibly more complicated, functions of the data.

This may motivate the claim that among all methods of estimation, the "straight rule" (use f_n as an estimate of f) is in a sense outstanding. In other words: without further information it seems straightforward to employ Ockham's razor and to use the least complicated estimation method. In a sense this is true, since the complexity of the straight rule is minimal relative to all other rules whose estimates *deviate* from the observed f_n. Whilst the straight rule corresponds to $f_n = g(f_n)$, i.e., g is the identity function, any other rule needs to specify a non-trivial and thus more complex function g. However, the whole point of estimation is to get an educated guess of some parameter (or property) of a population, *exceeding* the sample.

Without explicit reference to such a population it remains unclear if some method of estimation hits its target, be it straightforward or not. (In particular, there is no explicit criterion such as $|g(f_n) - f| < \epsilon$, since without a population, there is also no population parameter.) It may be easiest to project the current state into the future

[34] As always, these results should be applicable if the mechanism producing the data is sufficiently close to ideal coin tossing and the number of observations is rather large. In general, any formal account makes assumptions or imposes conditions. If these are met approximately in reality, it seems justified to apply the mathematical theory.

[35] Why should it - without any assumption? It suffices that the population is not "stable", in the sense that the proportion f is not constant.

without any modification, however, in general, such a heuristic does not seem to be a sophisticated method of estimation. Moreover, the basic problem remains that the gap between any finite vector and some limit, being a property of an infinite sequence, is unbounded. It does not help to keep silent about the population, i.e, the upper tier.

Owing to the theory of complexity, the remarkable phenomenon of "apparent convergence" can be explained *without* reference to a (real) limit: Virtually all finite binary strings have high Kolmogorov complexity, i.e., they are virtually non-compressible. According to Fine's theorem (cf. Li and Vitányi (2008), pp. 141), the fluctuations of the relative frequencies of these sequences are small, that is, $\max_{1\leq j\leq n} |f_j - f_n| < \varepsilon$ for all $\varepsilon > 0$. Both facts combined explain "why in a typical sequence produced by random coin throws 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).

To conclude these matters, it ought to be mentioned that formal learning theory finds itself in a quite similar situation: Given a certain knowledge base, the aim is to learn, i.e., to extend this base. Quite obviously, how far one can get crucially depends on the methods allowed or available. This establishes a close link with the theory of computability and proof theory. More generally, it is no surprise that there are always limits as to what is achievable: Given some resources (objects, concepts, boundary conditions, methods, etc.) the results obtainable with the help of these means will always be bounded in some way. And there is no free lunch - without substantial assumptions, hardly any nontrivial move can be vindicated.

5.6 Goodman's new riddle

Stalker (1992) gives a nice 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 given in section 3.3, 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. For example, suppose 'grue' applies to all things examined before t if and only if they are green and to other things if and only if they exist in a world in which pigs have wings. Then, if emeralds examined before t are green, we can predict that after t pigs will have wings. (ibid.)

In other words, instead of criticizing induction like so many successors of Hume, 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)? Since any experiment is made under very special conditions, it is not clear if results obtained by a particular observer can be generalized at all:

> If 'grue' is redefined to apply to all things examined in the observer's own laboratory if and only if they have one characteristic, and to other things if and only if they have a different one, then an analogous argument seems to 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.

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.

Looking at Illustration 3 gives us a clue 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 the following situation, where the uppermost line is unbounded:

[Illustration 4: 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. Considering GRUE: The concrete level consists of the constant colour green, i.e., a single point. The abstract level 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 to tomorrow, etc.) indicating when the change of colour happens, and to define $t = \infty$ if there is no change of colour. Clearly, the set $\mathbb{N} \cup \{\infty\}$ 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.[36] 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*.

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, summarizing the latter's ideas, it's the second "any" that does the harm: "any sizeable sample very probably matches its population in any specifiable respect" (cf. Williams (1947), p. 100). 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. What is true, however, is the statement that, given a certain or a 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 is perfectly correct when he concludes that a finite number of observations only allows one to distinguish between a finite number of hypotheses.

The second, often more severe, problem consists in finding the right 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, it may be non-existent or (almost) endless. The stance taken here 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 therefore always consists of the sets (objects), dimensions (properties) and boundary conditions involved, plus (at least) two levels of abstraction in each dimension.

6 Common structure of inductive problems

Throughout this contribution, the basic paradigm for an inductive problem is two levels of abstraction being connected in an asymmetric way - typically via a certain operation. The most primitive model used is thus two sets, connected by the subset relation $\subseteq$. However, in practice (or more realistically), there are many different kinds of operation and corresponding tiers. Here is a (non-exhaustive) list:

[36] 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.

Less information A (subset, lower tier)	Relation $\subseteq$	More information B (superset, upper tier)
Object (the "positive")	compatible with	Formula (form, the "negative")
Model	exemplify	Logical expression
Model	specify	Axiom system
Examples	satisfy	Constraints
Element	member of	Set
Physical object	comply	Physical law
Sample	draw	Population
Realization		Random variable & distribution
Data at hand		General hypothesis
0 or 1	replace	Wildcard
A string encoding a certain amount of information	compress	A shorter string encoding at least as much information
Qualitative description (imprecise, roundabout)	refine	Quantitative statement (precise, "information-loaded")
Data at hand	inform about	General hypothesis, theory
Single function	embed	Ensemble of functions
Single object		Set of similar objects
Pieces	assemble	Whole
Parts & interactions	constitute	System
Specific instant	generalize	General Law

For example, a formula defines a general form, and there may be concrete objects that comply with this form. Solving a mathematical problem in a narrow sense means that a formula (constraint, boundary condition), such as $a^2 + b^2 = c^2$ is given and has to be filled with concrete numbers. In a more general sense, one first has to find an appropriate formula(tion) for a given scenario. Then, the task is to look for concrete objects that fit into this form. The most important class of examples are differential equations ruling important parts of physics, like the wave or the heat equation; Maxwell's, Hamilton's, and Navier-Stokes equations, etc.

If the formula is a physical law, physical objects or phenomena have to satisfy it. That is, their behaviour is governed by the law, at least approximately. In mathematics, the basic paradigm is defined by some axiom system, and concrete models are observe them. More typically, in applications, there are certain boundary conditions (top-down restrictions), and the main problem consists in finding concrete solutions that satisfy them all. For example, a system of equations is given, and one is looking for a particular solution or - most ambitiously - one tries to find all possible solutions. In pure mathematics, one always considers particular objects (points, numbers, functions, etc.) and embeds them in a natural ensemble of similar objects (typically called "spaces").

In statistics, the upper tier is defined by some population having certain properties (e.g., a distribution or parameters). Samples provide the investigator with information about the population, for example, since the sample has been "drawn at random" from the population. (More technically speaking, in parametric statistics, observations $x_1, \ldots, x_n$ are an iid sample from the population $X \sim P_\theta$.) The task is to estimate properties of the population. In more general terms, there is data and one is interested in general hypotheses or laws. In this vein, (Fisher 1935/1966), p. 16, said:

> Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis.

In the information sciences, binary strings are the most concrete objects, and strings with some degree of freedom (e.g., wildcards) are their corresponding general counterparts. Given a slightly different perspective, a string of a certain length contains some amount of information and compressing it means to store the same amount of information in a more compact way. Another class of examples is given by a system and its parts. Given the pieces, one has to assemble the whole picture. More typically, there are modules and their interactions that have to be put together to a well-proportioned system.

The operations connecting two tiers are quite diverse. For example, *insert* a specific number, object or phenomenon that adheres to some formula, axiom system, boundary condition, or law. *Find*, *assemble*, or *construct* some solution (object) that observes all boundary conditions. *Define* a concrete model or *give* a specific example, *replace* some variable by a certain number, *consider* the realization of a random variable, concrete data, *make* a random experiment (e.g., toss a coin), *select* a typical case. *Calculate* some number using a given algorithm, *execute* a certain operation, instruction, or program, *apply* a general law to a specific instance. *Store* a certain bit of information in a less concise way, *proceed* from an exact to a less precise or less comprehensive description/explanation. *Reduce* or *decompose* a system into its components. What does a certain assumption *imply*? What are its *consistent* consequences?

In the context of discovery, typically either one of the tiers, or the operation connecting them appears first. For example, algebraic operations, but also differentiation or geometric transforms have been used for centuries, however, the set of objects to which they can be applied has been growing steadily. While a mathematical investigation often starts with a concrete situation and generalizes it to a reasonable extent, the reverse is true in technical applications: There, the main task rather consists in applying a general theory to the problem at hand.

It may further be mentioned that in the context of justification, a theory can most easily be developed if one starts with the more informative upper tier. Thus many approaches stress this constituent, and may quite easily overlook the role of other constituents and their interactions. For example, in the structuralist philosophy of mathematics (cf. Shapiro (1997)), a specific object is defined as a "certain position in a given structure". Quite similarly, in synergetics, the ensemble "enslaves" its parts, and in macroeconomics, fundamental parameters of the economy seem to be decisive for the success or failure of some enterprise.

Be this as it may: The basic reason why we speak of tiers and not of modules is the fact that the former do not interact on an equal footing. That is, sets A and B are connected in an *asymmetric* way. More specifically: It is always possible to proceed from the more abstract level (containing more information) to the more concrete situation (containing fewer 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 operations are *not* always possible or well-defined. They only exist sometimes, given certain additional conditions, in specific situations. Even if they exist,

it may be impossible to apply them 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, truth may not be provable, and it can be very difficult to find a causal relationship behind a cloud of correlations. The most interesting questions are thus: What do the inverse operations look like, and when can they be applied? Here are some examples:

(i) Given certain (empirical) phenomena, what is the general law governing their behaviour?

(ii) Given a set of examples, what is the general formula? What is the common core of an entire class of related problems?

(iii) What are the constraints (top-down restrictions), characterizing certain objects? (The point is to find constraints that define a particular object.)

(iv) Infer from the sample to the population, e.g., $P(Hypothesis|Data)$.

(v) Generalize from numbers to distributions.

(vi) Given some data, try to find an appropriate model.

(vii) Compress the string x as much as possible, without changing (losing any of) the information I in the string x

(viii) Elaborate and refine (switch from ordinary language to mathematics, replace qualitative statements by quantitative ones, measure with a higher degree of precision).

(ix) Go from qualitative verbal descriptions to exact quantitative rules.

(x) Give reasons for external validity (how much does a specific study bear on a general problem)?

(xi) What can be said about B, when all we know is a subset A of B?

(xii) When (under which conditions) does the reverse to $A \Rightarrow B$ hold?

(xiii) Inverse mathematical operations: Minus (-), division (:), integration. When are they possible?

(xiv) Reconstruct the system from its parts and their relationships, i.e., figuratively speaking, “solve a puzzle”.

(xv) Can some concept, line of argument or theory be applied in a different field? Is it possible to unify various approaches on a more abstract level?

As could be expected, there is no general answer, since the situations considered and the operations defined for these frameworks are very different. However, it is also quite obvious that there is a continuum of reversibility: One extreme consists in perfect reversibility, i.e., given some operation, its inverse is also always possible. The other extreme is complete non-reversibility, i.e., the operation cannot be inverted at all. 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. However, typically, an operation can be inverted given some special conditions. These can be nonexistent (any differentiable function can be integrated), mild (division can almost always be applied), restrictive (in general, a^b is only defined for positive a) or even impossible (given a string x, find the smallest string containing the same information). That is also true for interesting mathematical theorems. Very often, they state "forward" conditions subject to which $A \Rightarrow B$ holds. It is then straightforward to ask: "What are the 'backward' conditions under which $B \Rightarrow A$ holds?

The most straightforward mathematical model of an operation that simplifies matters (wastes information) is a non-injective mapping: Given two sets A, B and $f : A \to B$, there are elements $a_i \in A$ having the same image $b = f(a_i)$ in B. Therefore, typically, inverting some interesting but straightforward operation leads to a larger class of objects, i.e., this set of objects is more general (abstract). When the Greeks tried to invert multiplication, they had to invent broken numbers, thus leaving behind the familiar realm of whole numbers. It is also no coincidence that the set of integrable functions is much larger than the set of differentiable functions. Here is a physical example: "The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. [...] The constructivist hypothesis breaks down when confronted with the twin difficulties of scale and complexity" (Anderson (1972), p. 393). For example, it is one thing to reduce life to biochemistry, in particular to the most basic carriers of information, i.e., genes. However, it is a formidable task to start with a gene and to construct the protein it encodes, let alone to start with a genome and reconstruct the whole organism. Typically, in order to understand some phenomenon, it is necessary to analyze it in much detail. However, reductionism is wrong in assuming that this, in general, is sufficient for a synthesis.[37]

7 Conclusions

Of course, nitpicking is endless: What is an appropriate framework, shouldn't we consider different assumptions, isn't there a better alternative model, is Hume's problem (mainly) contingent or not? Due to the nature of induction (see the last section), a solution will always consist of a "core" of nice examples and a "penumbra" of rather pathological cases. Distinguishing them in a precise manner is science, philosophy's task can be understood to give a general account (like this contribution).[38]

The result of the latter analysis is unequivocal, the various models studied agree: Although there are outliers, as a rule, induction can be justified (given certain conditions, of course). That is, the core area of sound rational inductive reasoning is rather large, and the pathological boundary is rather narrow:

[37] Perhaps such cases, when invertibility is very difficult, lend credibility to Hume's otherwise amazing claim that only deduction can be rationally justified, or that induction does not exist at all (Popper).

[38] However, actual practice, at least during the last decades has been different, see Williams (1947), pp. 15-20, Stove (1986), pp. 86-88, 214-216, and Groarke (2009).

(i) Typically, iterative procedures lead to reasonable results (in particular, recursive algorithms terminate and series converge). In other words, the threat of pathological circularity (e.g., tautologies) is smaller than it may seem at first sight.

(ii) Statistical models (in particular, population & sample) are perhaps easiest to treat. They highlight that if information accrues it is difficult to avoid convergence toward a reasonable limit. It is no coincidence that Williams (1947) gave the first concrete example of a rationally justified induction, thus disproving Hume's general claim.[39]

(iii) More generally speaking, mathematical concepts and models give constructive answers in situations where purely verbal means have failed. For example, convergence may still prevail although it need not be fast or monotonous (e.g., the probability of a hypothesis H, given sequential evidence $e_1, e_2, \ldots$).

(iv) Alas, the counterexamples put forward in the philosophical discussion are either purely verbal or consist of extreme cases. In particular, the strict division between the land of Future and Past is an inadequate model, since it does not take into account that the boundary is constantly receding in favour of the observer. It also does not suffice to point at a potential difficulty, e.g., circularity, in order to show that induction is impossible. To be convincing, criticism should/must be as detailed as the arguments criticized.

(v) Perhaps the fallacy most often encountered is the desire to treat induction like deduction. More specifically, it is the idea that if there were a general inductive rule, particular inductive steps could be (deductively) justified, motivating "the great quest" for some general inductive principle. For the reasons given above, this is a rather hopeless endeavour. Moreover, on logical grounds, a general law is strong, since it - deductively - entails specific consequences. Alas, induction is the opposite of deduction, and therefore some general inductive principle (being the limit of particular inductive rules) would have to be *weaker* than any specific inductive step (see section 5.2.2 and Saint-Mont (2011), pp. 348).

In sum, formally as well as empirically, Hume's problem can be dealt with in a constructive way. It turns out that the key concept is information. If mutual information $I(A;B)$ is larger than zero, it is possible to say something about A given B (and vice versa). To this end, there has to be a link between A and B. In theoretical settings the logical link necessary is provided by some assumption, in practice by some causal or probabilistic connection. The link, e.g., between past and future, is crucial, but not some specific kind of linkage, e.g., "likeness". Hume was right in noticing that some strong condition like resemblance may easily be violated. However, in practice a weak kind of concatenation may hold and suffice for an educated guess.[40]

In theory (or in principle) almost any kind of coupling will do, *as long as the link transports some information.* Here is a straightforward setting: If information flows from A to B but not from B to A (e.g., future becoming past, or, equivalently, the past growing monotonically), B is able to learn something about A, but not vice versa. Arguably, the most straightforward way to model a linkage between two sets is the subset relation, i.e., $A \subseteq B$, or, almost equivalently, to consider a sample A *from* some

[39] A few philosophers like Stove (1986) and Campbell (2001) defended his solution.

[40] The basic fact that the present connects the lands of Past and Future is necessary but not sufficient.

population B. 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 $A \subseteq B$, the larger the ratio $|A|/|B|$, the more one knows about B, given the sample A. In particular, if $A = B \Rightarrow |A|/|B| = 1$, and there is no inductive problem; if $|A| = 0$, i.e., if there is no sample from B, nothing can be said about B. Enumerative induction is not convincing since in this case $|A|$ is a finite number, but $|B| = \infty$. However, as long as the ratio $|A|/|B|$ remains finite, it seems to be fair to say that there is some evidence in A about B, and the closer $|A|/|B|$ is to one, the stronger this evidence is.

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

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

(ii) Inductively, i.e., from A to B, leaping over the gap $I(B) - I(A)$

(iii) Additively, i.e., $I(A) + c = I(B)$, with $c \geq 0$ bridging the gap, see equation (3).

More generally speaking, a model should formalize $I_B(A)$, the information in A about some property or parameter of B. In statistics, A is represented by a sample and B is some population. Bad samples and difficult estimation problems have in common that $I_B(A)$ is low, although the size of the sample might be considerable. In the worst case, $I_A(B)$ is vanishingly small, despite the fact that $|A|$ and $|B|$ do not differ much.[41] This is where random sampling comes in: Due to combinatorial reasons, typically $I_B(A)$ is large, although the size of the sample may still be rather small. Of course, highly informative samples can also be attained in a systematic manner which is the core idea of purposeful sampling (Onwuegbuzie and Leech 2007). For example, when trying to find the highest tree on earth, it seems to be wise to focus on forests containing species of tall-growing trees.

Unlike these models, pinning down the crucial ingredients of the problem, it is quite typical for verbal discussions that they easily miss their target. The whole philosophical discussion seems to be revolving 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. (Equivalently, one can say that uncertainty (errors) grows very fast, soon exceeding any bound.)

[41] If every observation contributes (about) the same amount of information c, all samples of size n are equivalent in the sense that they contain (about) the same amount of information cn. In other words, in such "nice" situations, quite typical for classical statistics, there are neither good nor bad samples. 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.

This example also demonstrates that in our universe, *as a matter of fact*, predictions can be impossible - locally, given certain unfavourable conditions. Thus the basic contingent question becomes why "we are 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 possess, in Peirce's words, a "truth-producing virtue" (cf. Braithwaite (1953), p. 264)? It is also consistent to ask for "the minimum form of the inductive postulate which will, if true, validate accepted scientific inferences," as Russell (1948), p. 400 did.

The roundabout answer is that persisting patterns are the rule, i.e., that, typically, information decays rather slowly. If information is at least conserved to some extent, predictions are feasible. If we acknowledge that a certain (often lasting) form, object or pattern is a particular and particularly important expression of information, it is structure (which may be understood as preserved information) rather than conformity that is vital: One can have reasonable doubts in nature's uniformity, however, sophists aside, hardly anybody will challenge the fact that we live in a structured universe. Since life would be quite impossible if basic rules changed every other day or in a rather arbitrary manner, it is our very existence that provides considerable evidence for the hypothesis that at least the part of the universe we inhabit *has been* inductively normal for quite some time. In this sense, our (biological) existence implies that induction has worked in our vicinity.

In sum, inductive steps can often be justified. Because of the evidence that has accrued, an extreme position (no inductive leap is ever justified) no longer seems to be rationally supportable. The opposite is true: Given any reasonable framework, in particular the world we inhabit or appropriate mathematical models, inductive steps can be vindicated. Reality chooses to be with conservative, bounded induction rather than with principled doubt. B. Russell (1912), p. 14, was right when he thought that "... induction must have validity of some kind in some degree..."

References

Achinstein, P. (1961). The circularity of self-supporting inductive arguments. Analysis, 22(6), 138-141.

Anderson, P.W. (1972). More is different. Broken symmetry and the nature of the hierarchical structure of science. Science 177, 393-396.

Augustin, T.; Walter, G.; and F.P.A. Coolen (2014). Statistical inference. Chapter 7 in: Augustin, T.; Coolen, F.P.A.; de Cooman, G.; and M.C.M. Troffaes (eds.). Introduction to imprecise probabilities. Chicester: Wiley, 135-189.

Barnett, V. (1999). Comparative Statistical Inference, 3rd ed. New York: Wiley.

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. (1954). Problems of analysis. London: Routledge.

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

Bonjour, L. (2014). In defense of pure reason. Cambridge: Cambridge Univ. Press.

Bookstein, F.L. (2014). Measuring and Reasoning. Numerical Inference in the Sciences. New York: Cambridge Univ. Press.

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.

Brown, M.B. (1987). Review of Stove (1986). History and Philosophy of Logic 8, 116–120.

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.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *J. of the Royal Statistical Society A* **158(3)**, 419-466.

Cornfield, J.; and Tukey, J.W. (1956). Average Values of Mean Squares in Factorials. *Annals of Mathematical Statistics* **27**, 907-949.

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

Cox, D.R. (1986). Some General Aspects of the Theory of Statistics. *International Statistical Review* **54(2)**, 117-126.

Cox, D.R. (1990). Role of Models in Statistical Analysis. *Statistical Science* **5(2)**, 169-174.

Cox, D.R. (1995). The Relation between Theory and Application in Statistics (with discussion). *Test* **4(2)**, 207-261.

Cox, D.R. (2000). Comment on Lindley, D.V. (2000). Philosophy of Statistics (with discussion). *The Statistician* **49(3)**, 293-337.

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

Edwards, P. (1974). Russell's doubts about induction. Chap. 2 in Swinburne (1974), 26-47. Warming. *The MIT Press.*

Edwards, P.N. (2010). A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming. *The MIT Press.*

Efron, B. (1978). Controversies in the Foundations of Statistics. *American Math. Monthly* **85(4)**, 232-246.

Efron, B. (1998). R. A. Fisher in the 21st Century: Invited paper presented at the 1996 R. A. Fisher Lecture. *Statistical Science* **13(2)**,95–122.

Einstein, A. (1954). Ideas and opinions. New York: Crown Publishers.

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.

Fisher, R.A. (2003). Statistical methods, experimental design, and scientific inference : a re-issue of statistical methods for research workers, the design of experiments, and statistical methods and scientific inference. Oxford : Oxford Univ. Press.

Freedman, D.A. (2010). Statistical Models and Causal Inference. A Dialogue with the Social Sciences. New York: Cambridge University Press.

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

Giaquinto, M. (1987). Review of Stove (1986). Philosophy of Science 54, 612–615.

Giere, R.N. (1999). Using Models to Represent Reality. In: Magnani et al. (ed.) Model-Based Reasoning in Scientific Discovery. *Springer*.

Greenland, S. (1990). Randomization, Statistics, and Causal Inference. *Epidemiology* **1(6)**, 421-429.

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.

Gustason, W. (1994). Reasoning from evidence. Inductive Logic. New York: Macmilan.

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

Hampel, F.R. (2005). The Proper Fiducial Argument. *Electronic Notes in Discrete Mathematics* **21**, 297-300.

Howson, C. (2000). Hume's problem: Induction and the justification of belief. Oxford: Clarendon Press.

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

Hume, D. (2010). 1st ed. 1740. An Abstract of a Book Lately Published; Entituled, a Treatise of Human Nature. Gale Ecco, Print Editions.

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

Hutter, M. (2007). On Universal Prediction and Bayesian Confirmation. *Theoretical Computer Science*, **384**, 33-48.

IEP staff (2015). Deductive and Inductive Arguments. Internet Encyclopedia of Philosophy (Sep. 9, 2015, see archive.org)

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.

Jeffreys, H. (1973). Scientific Inference. (3rd ed.) 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.

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

Lehmann, E.L. (1990). Model Specification: The Views of Fisher and Neyman, and later Developments. *Statistical Science* **5(2)**, 160-168.

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

Lipton, P. (2004). Inference to the Best Explanation (2nd ed.). London: Routledge.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychol. Bull.* **70**, 151-159.

Maher, P. (1996). The hole in the ground of induction. Australasian J. of Phil. **74**, 423-432.

Maskin, E.; and A. Sen (2014). The Arrow Impossibility Theorem. New York: Columbia Univ. Press.

Meehl, P.E. (1990). Apraising and amending theories: The strategy of Lakatosian defence and two principles that warrant it. *Psychological Inquiry* **1(2)**, 108-141.

Mill, J.S. (1843). A System of Logic, Ratiocinative and Inductive. London. Cited according to the 1859 ed., New York: Harper & Brothers.

Miller, I; and M. Miller (1994). Statistical Methods for Quality: With Applications to Engineering and Management. *Prentice Hall, Upper Saddle River, NJ.*

Miranda, E.; and G. de Cooman (2014). Lower previsions. Chapter 2 in: Augustin, T.; Coolen, F.P.A.; de Cooman, G.; and M.C.M. Troffaes (eds.). Introduction to imprecise probabilities. Chicester: Wiley, 28-55.

Morgan, M.S.; and M. Morrision (eds., 1999). Models as mediators: Perspectives on natural and social sciences. New York: Cambridge University Press.

Onwuegbuzie, A.J.; and N.L. Leech (2007). A call for qualitative power analyses. *Quality & Quantity* **41**, 105-121.

Papineau, D. (1992). Reliabilism, induction and scepticism. *The philosophical quarterly* **42(166)**, 1-20.

Pearl, J. (2009). Causality. Models, Reasoning and Inference. (2nd ed.) *Cambridge University Press.*

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

Pitman, E.J.G. (1957). Statistics and Science. *J. of the American Statistical Association* **52**, 322-330.

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

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.

Royall, R.M. (2000). On the Probability of Observing Misleading Statistical Evidence (with discussion). *J. of the American Statistical Association* **95**, 760-780.

Russell, B. (1912). The problems of philosophy. London: Home University Library.

Russell, B. (1948). Human knowledge: its scope and limits. London: Home University Library.

Saint-Mont, U. (2011). Statistik im Forschungsprozess. Heidelberg: Physika.

Shadish, W.R.; Cook, T.D.; and D.T. Campbell (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company.

Shapiro, S. (1997). Philosophy of Mathematics. *Oxford University Press, New York, Oxford.*

Skyrms, B. (2000). Choice & Chance. An introduction to inductive logic (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.

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.

Steen, L.A.; and J.A. Seebach (1995). Counterexamples in topology. Dover Publ.

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.

Van Cleve, J. (1984). Reliability, justification, and the problem of induction. In: French, P.; Uehling, T.; and H. Wettstein (eds.) Midwest Studies in Philosophy 10, Minneapolis: Univ. of Minnesota Press.

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.

Wallace, C.S. (2005). Statistical and Inductive Inference by Minimum Message Length. New York: Springer.

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.