Synthese (2023) 201:184
https://doi.org/10.1007/s11229-023-04172-9

ORIGINAL RESEARCH

Inductive knowledge under dominance

Marco C. Campi[1]

Received: 18 October 2021 / Accepted: 24 April 2023 / Published online: 16 May 2023

Abstract
Inductive reasoning aims at constructing rules and models of general applicability from a restricted set of observations. Induction is a keystone in natural sciences, and it influences diverse application fields such as engineering, medicine and economics. More generally, induction plays a major role in the way humans learn and operate in their everyday life. The level of reliability that a model achieves depends on how informative the observations are relative to the flexibility of the process by which the model is constructed. When the process is articulated so that the model can incorporate descriptive details and subtleties, a large set of informative observations are required to reliably tune the model, whereas models obtained from simple procedures can be tuned with fewer observations. This article introduces the concept of "dominance", which refers to the situation in which a reduced subset of observations suffices to reconstruct the model. A mathematical framework is presented to quantify the reliability of learning procedures as a function of the size of the subset of dominant observations. Although limited in scope, we believe that our study can contribute to the understanding of some fundamental mechanisms by which knowledge is generated from observations in inductive reasoning.

Keywords Inductive methods · Dominance · Generalization · Quantitative reliability

1 Introduction

Inductive reasoning refers to the process of synthesizing general principles and descriptive schemes from observations. This is key not only to all physical sciences, it also proves fundamental to large areas of applied fields in engineering, medicine and economics. Even more broadly, induction forms the footing of all processes by which we learn from experience, with paramount implications in the endeavor of describing the world we live in and of making decisions on how to operate on it. But, in virtue of

✉ Marco C. Campi
marco.campi@unibs.it

[1] Center for the Study of Inductive Methods c/o Department of Information Engineering, University of Brescia, via Branze 38, 25123 Brescia, Italy

what can observations be used to derive principles and to construct schemes meant to be applied to new, out-of-sample, cases? Why can observations drive our way of thinking and acting in situations that we have not previously encountered? While these questions have been debated for at least four centuries, in this article we aim to suggest a new, mathematically-structured, framework that we believe can shed new light on this topic. To this purpose, after informally introducing in this section some general concepts and ideas, we will move to a more formal presentation in the next Sect. 2 that culminates in results stated in the form of propositions.

To describe knowledge, we use a *model*. A model is a descriptor of a *population*, where the word "population" must be thought of in broad sense, it can be a collection of individuals, but it can also refer to a group of objects or even to a set of conditions in which a device operates. See Sect. 2 for examples. After the model has been built and validated, it provides an instrument of investigation, Morgan and Morrison (1999), and it allows for *surrogative reasoning*, Swoyer (1991) and Contessa (2007). In science, models are used in prediction and to support decisions, a fact that has been widely discussed in the literature, see e.g. Bailer-Jones (2009), Frigg and Hartmann (2012), Hughes (1997), Magnani et al. (1999), and Magnani and Nersessian (2002).

In this study, we are interested in models that describe specific *attributes* of a population, Suppes (1960) and Da Costa and French (2000). The attributes recognize and isolate a small number of salient characteristics, Maki (1994), so that the model incorporates a high level of *idealization*, Cartwright (1989, Ch. 5). In more specific terms, the models we consider are *subsets of the domain in which the attributes are defined* (such models are also called "set models"). A set model is meant to identify a portion of the domain that captures the variability of the attributes in the population of interest, that is, the model covers the range of the values that characterize the population. For example, in a demographic study a set model can describe the variability of wealth in a population and, in a medical application, the set model can delimit the values of the outcomes of a clinical test given to a population that suffers from a certain disease.[1]

In normal cases, the population is too large for us to access one by one all of its members. Hence, models are constructed from a reduced number of observed instances of the population called *observations*: the attributes of a *sample* of observations drawn from the population are recorded and used to derive a model meant to describe the whole population. In other words, one constructs what is called an *observation-driven model*, van Fraassen (1980), Suppes (1962), Laymon (1982), Woodwart (1989), Mayo (1996), McAllister (1997), and Harris (2003). In science, this is how demography and social sciences operate, it represents a fundamental framework for fields like medicine, engineering and finance, and even laws of physics are constructed using this approach.[2]

[1] It is important to observe that not all existing models operate in the way described here. For example, in some applications one wants to construct a line of best fit. We shall describe alternatives, so as to better position our contribution, in Sect. 2 after we introduce a formal definition of set model. We also advise the reader that, throughout this article, when we use the word "model" this will stand for "set model" unless otherwise specified.

[2] For example, social demography studies the relationships between economic, social and cultural features of a society from the analysis of a sample elicited from the population; the penetration of machine learning techniques aiming at constructing classifiers from an observed set of cases (the so-called *training sequence*) is getting ever more pervasive in medical diagnoses as well as in control, telecommunications and computer

Two principles drive, in one way or another, the process of learning a set model: (i) the model must accommodate, and correctly describe, the available sample obtained from the population; and (ii) the model should provide a tight coverage of the population (that is, cases that are not compatible with the population should be left out of the model) so as to make the model informative and useful. How these principles can be formalized and put in practice in specific procedures is discussed in subsequent sections.

Since the model is constructed from a restricted sample of observations, it cannot be expected to be an exact descriptor of the whole population, and it will bear a certain degree of *imprecision*. As a consequence, when applied to a new member of the population we have to consider that the model can err and return an incorrect evaluation. A desirable feature is that the model errs rarely, a fact that is at times expressed by saying that it has a good *generalization capability* and in this article we study the elements that concur in determining the generalization capability of observation-driven models. Fundamental questions that we shall address are:

- can a moderate number of observations be used to describe a large population?
- what does the generalization capability depend on?
- are there universal methods to exert control over the generalization capability?
- and, even more fundamentally, why can knowledge that originates from a limited set of observations be applied to new, yet unseen, cases?

While this last question touches what may appear to be an unjustified stretch, it is plain that it describes the way science operates all the time: science does not provide look-up tables of previous examples, it aims at generalization and prediction,[3]

In previous work in the context of curve and causal model fitting, Akaike (1974) has proposed a criterion to evaluate the predictive accuracy of models tuned on a set of observations. Akaike's approach, however, introduces the restrictive assumption that a candidate model exactly describes the population. This assumption has been somehow relieved by the TIC criterion, Takeuchi (1976). See Forster and Sober (1994), Sober (2008), Forster and Sober (2011), and Sober (2015) for more-in-depth discussions on these methods. The background of knowledge this article builds upon lies in *statistical learning*.[4] In specific terms, in this article we consider learning schemes in which some observations are more important than others to determine the model, a property that we call *dominance*, see Definition 1 in Sect. 2 for a precise definition of *dominant observation*. We show that dominance applies broadly to inductive processes. The number of dominant observations is called *complexity* (see again Definition 1, along

engineering; and, certainly, predictors built from previous measurements (e.g., rates-of-return) are broadly used in quantitative finance. Also in physics laws are built by generalizing from a limited number of observed cases; for example, electrons are deemed to have negative charge because all electrons that have been thus far tested in a laboratory had this property (this is an example of *enumerative induction*, Example 3 provides another example of this type).

[3] Bruno De Finetti, referring to Henri Poicarè wrote in de Finetti (1989) "*he has clearly understood that only an accomplished fact is certain, that science cannot limit itself to theorizing about accomplished facts but must forsee, that science is not certain.*"

[4] More precisely, in a sub-branch of statistical learning that aims to establish the coverage of set models. More generally, statistical learning studies how well models of various nature describe a population, for example according to a criterion of average fit.

with Definition 2). We then prove that the *reliability* of a learning procedure (precisely defined in Definition 4) can be evaluated from its complexity. This result has profound philosophical implications. Indeed, a direct evaluation of the reliability of a learning procedure would require to test how models constructed using the procedure perform on new, yet unseen, cases. In turn, this would require to have a full description of the population, under which circumstance the problem inductive reasoning deals with would disappear altogether. In contrast, the complexity only depends on the procedure and can therefore be determined without any knowledge on the population. Hence, from the results stated later as Propositions 1 and 2 (these propositions link reliability to complexity) one secures a theoretical ground to ascertain the reliability of learning procedures without requiring any prior knowledge on the underlying population from which observations are obtained. These results offer a rational explanation of the mechanisms by which trust in inductive learning is generated and carries profound implications on the possibility of acquiring knowledge from examples.

The structure of this article is as follows. In the next section, the notions of *dominance*, *complexity* and *reliability* are introduced and it is shown that a precise relation links the degree of reliability of a learning procedure to its complexity. One can go beyond complexity of a learning procedure and speak of complexity of the model generated by the learning procedure. This has implications on the way the procedure operates and a procedure that accepts a model only when its complexity is moderate provides a higher degree of reliability. This approach is analyzed in Sect. 3. In Sect. 4 we draw conclusions and discuss additional implications of the achievements of this article in relation to Popper's refutation theory. All formal derivations of the results are provided in Sect. 5, which can be skipped at first reading.

2 Complexity and reliability

In this section, we formalize the concept of model used in this article and that of a learning procedure that constructs models from observations, and then introduce a definition of complexity of a procedure and show how the reliability of the procedure can be evaluated from its complexity. All results are valid irrespective of any special characteristic of the population, which means that they can be applied without requiring any prior knowledge of the population.

2.1 Models of attributes

Suppose that we are interested in given attributes of a population, and the goal is to build a description of these attributes using a (possibly moderate) sample of members taken from the population. A couple of very simple examples are useful to concretely illustrate the idea.

Example 1 (height) Suppose we are interested in the height of Italians, so that Italians is our reference population and the height of Italians is the attribute we want to describe. To judge how tall Italians are, we draw a sample of N Italians and measure their height. Then, we consider the tallest in the sample and construct a model that predicts that a

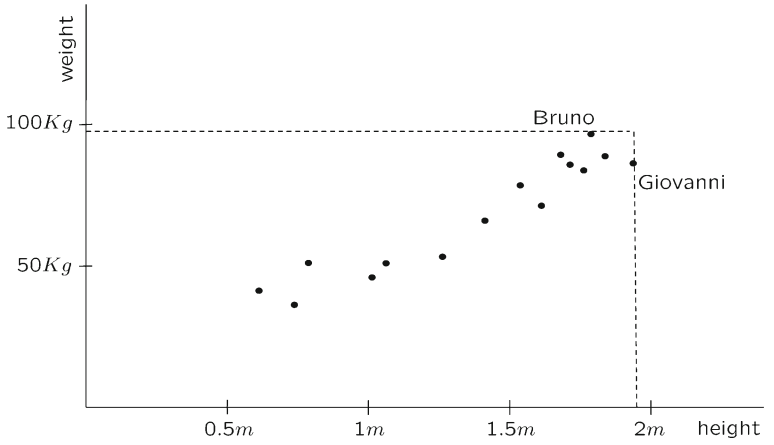

Fig. 1 A rectangle describing the height and the weight of Italians. Giovanni is the tallest in the sample and Bruno is the heaviest. A new Italian is predicted to be not taller than Giovanni and not heavier than Bruno

new Italian will not be taller than the tallest in the sample; this amounts to building an interval model expressed as $[0, \max\{height_i\}]$, where $height_i$, $i = 1, \ldots, N$, is the height of the i-th Italian in the sample. While this is a reasonable model, we certainly cannot expect it to be infallible. *

Example 2 (height and weight) Suppose that, besides the height, we are also interested in the weight of Italians. The height and the weight of N Italians are measured, so that this time each Italian is described by two numbers and an Italian corresponds to a point on a plane whose coordinates are height and weight, respectively. In this case we can use a rectangular model $[0, \max\{height_i\}] \times [0, \max\{weight_i\}]$ and predict that one new Italian will correspond to a point in this rectangle, that is, will not be taller than the tallest in the sample and not heavier than the heaviest in the sample, see Fig. 1.

An alternative to the rectangle is the model shown in Fig. 2. Here, the model is the smallest strip that contains the points corresponding to the Italians in the sample. This is called a minimal regression model, or a "Tchebyshev layer", Harter (1982) and Birkes and Dodge (1993).

Regression models are widely used in applied fields. Indeed, it is not rare that an attribute for which we have a special interest is difficult to measure and, hence, it is estimated from other, more accessible, attributes. Practical examples include medical applications where the health of a patient is estimated from the outcome of a clinical test. To understand in more detail how this scheme operates, suppose that the model in Fig. 2 is in use and that we have measured the height (here playing the role of the accessible attribute) of a new Italian to be 1.80 ms. Using the model in conjunction with the knowledge that the new Italian is 1.80 ms tall restricts our prediction to the intersection of the vertical line corresponding to 1.80 ms of height with the layer, yielding the line segment marked in bold in the figure, corresponding to the interval of

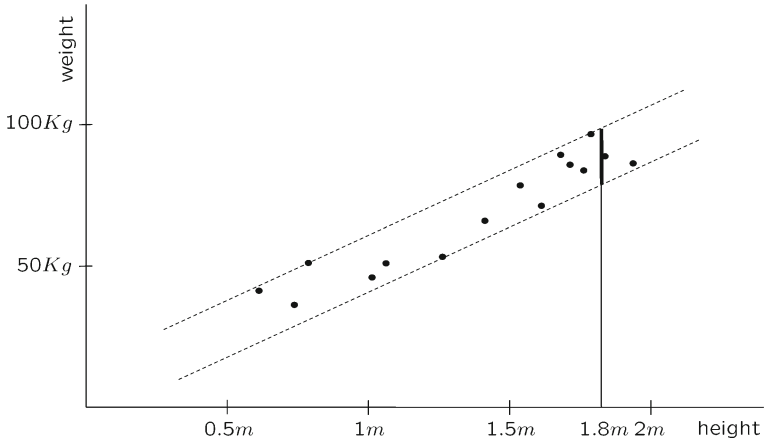

Fig. 2 A Tchebyshev layer is an instrument to regress one attribute of a population against other attributes. In the figure, the weight of Italians is regressed against their height

weights [76 Kg, 97 Kg] (this is the prediction of the attribute that is difficult to directly measure; in a medical application this is, e.g., the degree of severeness of a disease). In other words, the prediction is made by "cutting" the model corresponding to the observed attribute, or, more formally, if the value $\bar{x}'$ of attribute x' has been observed, the prediction is $\{x'' : (\bar{x}', x'') \in \text{ model}\}$, which are the values of the attribute x'' for which the couple $(\bar{x}', x'')$ is in the model. ∗

Generalizing from the previous examples, a population is described by various, say d, attributes. An attribute can be given by a number, as is in the case of the height and the weight, but, more generally, it can take values in any set, not necessarily a set of numbers. For instance, we can consider the color of the eyes as an attribute, and it takes value in the set $\{brown, blue, green, grey\}$. The d attributes of a member of the population are grouped in a list $x = (x', \ldots, x^d)$, and, when it does not generate confusion, we refer to x itself as a "member of the population", which means that we identify a member of the population with the list of its attributes. By $\mathcal{X}$ we denote the set of all potential values of x. If, e.g., the Italian population is described by the height and the color of eyes, then $\mathcal{X} = \{(x', x''); x' \in \mathbb{R}^+, \text{ the set of positive real numbers}, x'' \in \{brown, blue, green, grey\}\}$. The members in the sample are written as x_i, $i = 1, \ldots, N$, and the goal is to use the sample to construct a set M of instances of $x \in \mathcal{X}$, called a *set model*, which is meant to describe the population the sample is taken from.

2.2 A prototypical procedure for constructing set models

In this section, we describe a procedure by which set models are constructed from observations. Later, we shall discuss its applicability and use in real contexts.

In many modeling endeavors, a model M is selected from a pre-specified class of candidate models $\mathcal{M}$. Two requisites play an important role in the selection process, as described in what follows:

(a) M correctly describes the available sample;
(b) M does not introduce redundancy.

Requisite (b) expresses a principle of parsimony so that the prediction provided by the model is informative and useful; as we shall see, in formal terms this requisite sets a criterion of optimality. Instead, requisite (a) enforces that the model contains the members of the population that have been seen, and it sets constraints on the optimization process.[5] Before moving to a formal definition of procedure, let us review our height and weight example to facilitate an intuitive understanding.

Example 2—Cont'ed When a rectangle is used to describe the height and the weight of the Italian population as is done in the first part of Example 2, requisite (b) corresponds to requiring that the area of M is minimized; requisite (a) instead prescribes that the points in the sample are inside the rectangle. When considering a Tchebyshev layer (second part of Example 2), (b) prescribes that the layer's width be minimized under again the same constraint (a) that the points in the sample are inside the model. ∗

Setting a suitable criterion (b) of optimality is problem-dependent and, in a given application, the selection is based on reasons of convenience, also in the light of the intended use of the model. Irrespective of its choice, a criterion of optimality (b), along with a class of candidate models $\mathcal{M}$ and the constraints enforced by (a), define *procedure* P according to which M is selected.

Procedure P

- input: sample S;
- minimize with respect to $M \in \mathcal{M}$ the criterion (b)
  subject to condition: $x_i \in M$, for any $x_i \in S$;
- output: M that solves the optimization program in the previous point. ∗

Hence, M is the output of procedure P applied to the sample S, which justifies our using at times the notation $P(S)$ for M when we want to explicitly indicate the sample S that has been used.

Notice also that, for a given criterion (b), changing the class of candidate models and considering an enlarged class $\mathcal{M}' \supseteq \mathcal{M}$ ($\mathcal{M}'$ includes $\mathcal{M}$) improves, or, at worst, leaves unaltered the value of the optimality criterion (b) since the model that is optimal in $\mathcal{M}$ also belongs to the enlarged class $\mathcal{M}'$ and can therefore be selected when $\mathcal{M}'$ substitutes $\mathcal{M}$. This is illustrated by making again reference to our height and weight example.

Example 2—Cont'ed Suppose we select M from the class of all rectangles with sides parallel to the coordinate axes, instead of the class of rectangles with sides parallel to the coordinate axes and a vertex in $(0, 0)$ as it was done in Example 2. Then, the model M of smallest area containing the sample is $[\min\{height_i\}, \max\{height_i\}] \times [\min\{weight_i\}, \max\{weight_i\}]$, see Fig. 3. This rectangle is smaller, and hence more informative, than $[0, \max\{height_i\}] \times [0, \max\{weight_i\}]$. ∗

[5] In some cases, the model is allowed to fail on specific members of the sample that have a "odd" behavior as compared to other members (outliers) so that a smaller model, with improved descriptive capabilities, is achieved.

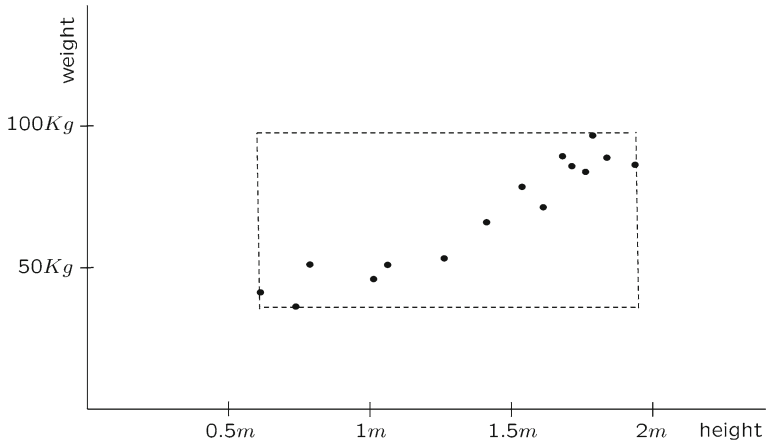

Fig. 3 The rectangle with sides parallel to the coordinate axes of smallest area containing a sample of Italians

Remark 1 It is important to notice that not all procedures aim at building set models, as is the case for the procedure P of this section. For example, *least squares* can be used to determine the "center of mass" of a given set of observations, which is a point-wise descriptor of the population. When observations are formed by two components (as it was in our regression Example 2), which we call here "input" u and "output" y respectively, a parameterized line—or, more generally, any parameterized curve, e.g., a polynomial—can be tuned to the observations by an average criterion of best fit. When this criterion corresponds to minimizing the sum of the squared errors between the measured outputs y_i and the values given by the parameterized line corresponding to the measured inputs u_i, one again speaks of *least squares*. Other criteria of best fit includes *total least squares* and the minimization of various functions of the error, like the exponential function in *risk-sensitive approaches*. We further notice that averaging procedures usually have infinite complexity with no dominant proper subsets; therefore, the analysis of this article does not apply to averaging procedures. ∗

In the remainder of this article we make reference to the prototypical procedure of this section. While this procedure is built on principles that govern many inductive constructions, still, as noticed in the previous remark, its coverage of inductive methods is partial and, hence, this article has a limited scope. We provide our results in the hope that the concepts presented here will contribute to the discussion on a mathematical understanding of inductive methods and will stimulate others to continue on this line of research.

2.3 Complexity of a procedure

The central issue of attention of this article is discussing the reliability of learning procedures. When the rectangle of Fig. 3 is used, the next individual is incorrectly predicted when s/he happens to be taller than the tallest in the sample or shorter than the shortest or heavier than the heaviest or lighter than the lightest. Clearly, the chance

of an incorrect prediction with this model is higher than with considering the rectangle in Fig. 1. What lesson can we learn from this? Intuitively, selecting from a larger class $\mathcal{M}'$ leaves us with more freedom to adapt to the sample of observations; as the model is steered towards a detailed description of the sample, it loses a grasp on the rest of the population. Said differently, with an enlarged class, the model has more capability to adhere to the sample at the expense of its generalization capability. In the height and weight example, at the extreme when a model can be any finite collection of points in $\mathbb{R}^2$, M can be selected to coincide with the sample itself and this model has no predictive capabilities on unseen members of the population.

To formalize this intuitive thinking, let us start by considering in more detail the idea that a model "adheres to the sample". In the height-and-weight example with M of the form $[0, \max\{height_i\}] \times [0, \max\{weight_i\}]$ as in Fig. 1, if Giovanni is the tallest and Bruno is the heaviest, then the constructed model "touches" the points in $\mathbb{R}^2$ representing Giovanni and Bruno, whose presence impedes the model from being smaller. One observation is now key: if someone were given only the two points representing Giovanni and Bruno, then this person would be able to reconstruct the model without seeing the other individuals in the sample by applying to Giovanni and Bruno the very same procedure according to which the model was constructed from the whole sample of N individuals. In this example, we say that Giovanni and Bruno are the *dominant* observations. The number of dominant observations is intuitively related to reliability issues: if a model can be seen as generated by few members in the sample and all other members in the sample agree with the model generated by these few, then these other members confirm the model and attest its reliability.

Interestingly, the number of members that generate the model using a given procedure is not always the same, it depends on the sample. For instance, in the height and weight Example 2 if Carlo had happened to be the tallest and the heaviest at the same time, then Carlo alone would have sufficed to reconstruct the model using the procedure.

The above concepts are formalized in the following definitions.

Definition 1 (*Complexity of a procedure for a given sample*) Given a sample S, the complexity of a procedure P in relation to S, written $c(P, S)$, is the smallest integer n such that there exists a sub-sample of n members in the sample so that the procedure applied to this sub-sample returns the same model as when the procedure is applied to the whole sample. The n members in the sub-sample are called the *dominant observations*. ∗

Definition 2 (*Complexity of a procedure*) The complexity of a procedure P, written $c(P)$, is the largest possible complexity of the procedure for a sample S, i.e., $c(P) = \sup_S c(P, S)$ where S is any finite subset of $\mathcal{X}$.[6] ∗

In the height and weight example with models of the type $[0, \max\{height_i\}] \times [0, \max\{weight_i\}]$, as we have seen above, the complexity of the procedure for a

[6] While complexity is one of the most debated and controversial concepts in science, and indeed it has attracted the attention of eminent mathematicians including Kolmogorov and Chaitin, Kolmogorov (1965), Chaitin (1966), Kolmogorov (1968), Definitions 1 and 2 do not want to contribute this discussion, they merely introduce measures of complexity within the specific setup here described of constructing models from observations.

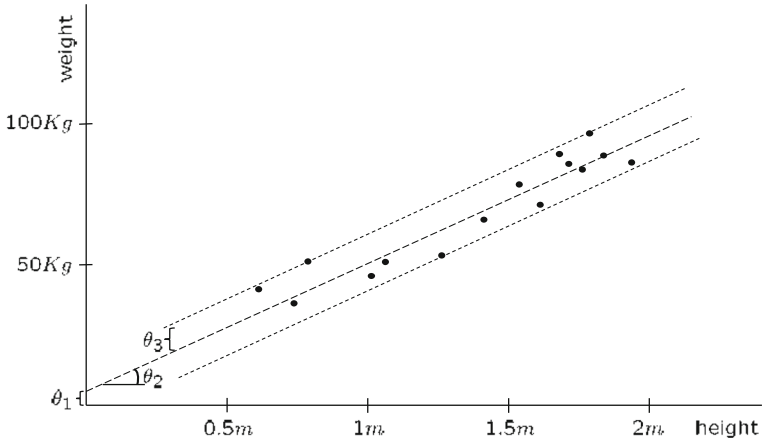

Fig. 4 Interpretation of the three parameters, θ_i, $i = 1, 2, 3$, that define a Tchebyshev layer

given sample can be 1 or 2, but it never exceeds 2. Hence, the complexity of this procedure is 2. Likewise, the complexity of the procedure that constructs a model of the type $[\min\{height_i\}, \max\{height_i\}] \times [\min\{weight_i\}, \max\{weight_i\}]$ is 4. For the determination of the complexity of procedures based on convex optimization one can resort to a noteworthy result proven in Calafiore and Campi (2005), Theorem 2, which states that a procedure with m optimization variables has complexity no larger than m. This result can, e.g., be applied to the Tchebyshev layer in the height and weight Example 2 as shown below.

Example 2—Cont'ed (height and weight - complexity when using Tchebyshev layers) The smallest strip that contains the points corresponding to a sample of Italians can be constructed by the following optimization program:

$$\begin{aligned} &\min_{\theta_1,\theta_2,\theta_3} \ \theta_3 \\ &\text{subject to: } |weight_i - [\theta_1 + \theta_2 \cdot height_i]| \leq \theta_3, \quad i = 1, \ldots, N. \end{aligned} \tag{1}$$

Indeed, for given values of θ_1 and θ_2, relation $weight = \theta_1 + \theta_2 \cdot height$ defines a straight line in the $(height, weight)$ domain, called the "central line" of the model. A visualization is provided in Fig. 4 where $\theta_1 = 6Kg$ and $\theta_2 = 52Kg/m$. Quantity $weight_i - [\theta_1 + \theta_2 \cdot height_i]$ is the vertical displacement of the weight of the i-th individual in the sample from the value taken by the central line corresponding to the height of this individual. The optimization program selects values for θ_1 and θ_2 so that an upper bound θ_3 on the maximum displacement in the sample, taken always positive thanks to the absolute value $|\cdot|$, is minimized. Hence, by program 1 all the individuals are "squeezed" into a layer that has the smallest possible width.

Program 1 is convex. Indeed, its objective function, θ_3, is linear, while the constraints $|weight_i - [\theta_1 + \theta_2 \cdot height_i]| \leq \theta_3$ can be rewritten by breaking the absolute value in its positive and negative part as follows: $-\theta_3 \leq weight_i - [\theta_1 + \theta_2 \cdot height_i] \leq$

θ_3, resulting in two linear inequalities in θ_i, $i = 1, 2, 3$. Each inequality defines a half-space in the $(\theta_1, \theta_2, \theta_3)$ domain and the simultaneous verification of the two inequalities holds in the intersection of two half-spaces, which is a convex set. Applying Theorem 2 in Calafiore and Campi (2005), we therefore conclude that the complexity of the procedure that constructs the Tchebyshev layer is no more than 3, the number of optimization variables in the problem.[7] ∗

In more general situations, computing the complexity of a procedure may not be an easy task. However, in view of the use of the notion of complexity in Proposition 1, overestimating the complexity still leads to a valid upper bound on the reliability (at the price that the upper bound can become somehow conservative). This observation can be used to alleviate the difficulty inherent in finding the exact value of the complexity.[8] Moreover, Proposition 2 in Sect. 3 applies to a context in which the reliability is only evaluated when the complexity for the sample at hand (i.e., the one actually used to construct the model) turns out to be not too high, in which case computing the complexity of the procedure is not required altogether.

Apart from practical aspects relating to its computation, what is central for this article is that the complexity is a property of the procedure. As such, it can be computed without relying on *a priori* knowledge on the population under consideration (*agnostic setup*). Hence, the certificates of reliability provided by the results in Propositions 1 and 2 by the only use of the complexity justify at a deep conceptual level inductive reasoning as a tool to construct models to predict the attributes of members of a population that have not been previously encountered. For example, considering that the procedure to construct Tchebyshev layers as in Example 2 has complexity 3, Proposition 1 allows one to draw quite strong reliability conclusions that do not depend on any prior knowledge, or conjecture, on how the population distributes on the 2-dimensional plane (refer also to the continuation of Example 2 after Propositions 1, where the height and weight example is resumed with numerical evaluations).

Remark 2 It may be surprising that the procedure for constructing a Tchebyshev layer has lower complexity than that for building the smallest rectangle with sides parallel to the coordinate axes, especially because the former is expected to produce tighter models than the latter. The reason for this expectation is that a Tchebyshev layer aptly incorporates the idea that weights and heights are correlated quantities, which is not possible with a rectangle. From this, we also see that domain knowledge does play a role to obtain better results in modeling procedures.[9] On the other hand, it is key that the rigorous validity of the results in the next Sect. 2.4 do not depend on whether or not the domain knowledge used in the problem formulation is accurate, or even if it is correct at all, they remain intact under all circumstances. This is the ground

[7] In fact, it is not difficult to show that the complexity of this procedure is exactly 3.

[8] For example, in Campi et al. (2018) viable approaches are provided to upper bound the "complexity of a procedure for a given sample" (Definition 1) based on the progressive removal of observations until no observation can be further removed without altering the model.

[9] Beyond this simple example, in modern decision-making problems dealing with complex systems, besides observations one does want to exploit domain knowledge that comes from various sources, often including some knowledge that, while not completely trustworthy, can still be of help to obtain a satisfactory model.

on which inductive knowledge finds its foundations in the face of the criticism that domain knowledge is as much in need of justification as induction itself. ∗

A last remark provides an additional interpretation of the concept of complexity. Informally, the complexity of a procedure can be thought of as a means to rank the complexity of the question that the procedure answers. For example, the procedure that constructs the model $[0, \max\{height_i\}] \times [0, \max\{weight_i\}]$ answers the question: what is the tightest model in agreement with the sample to predict how tall and how heavy a next individual can at most be? Similarly, the procedure that constructs the Tchebyshev layer answers the question: what is the tightest model in agreement with the sample to regress the weight of individuals against their height? Since the complexity of the latter procedure is 3 and that of the former is 2, we can think of the second question as being more complex than the first.

2.4 Reliability of a procedure

Let us now go back to the focus of this article: the reliability of inductive methods to construct models. We want to link the reliability of a procedure P to its complexity. Two Fundamental issues of investigation are:

(I) how does the reliability of a procedure depend on its complexity?

We expect that asking simple questions, viz. the procedure has low complexity, produces reliable models in such a way that the answer we obtain warrants our trust. How can this intuition be put on solid quantitative grounds?

Suppose next that we ask a difficult question, but we accept the answer only when the answer turns out to be simple, that is, the procedure has high complexity, but we use the model only when the procedure for the sample at hand has low enough complexity.

(II) Can we then be as confident in the use of the model as when a simple question is asked in the first place?

We want to make the above questions precise and provide answers supported by quantitative methods. The first step along this route is to formalize the concept of reliability.

When the model is obtained from a sample, it is natural to define the reliability of the model as the proportion of the rest of the population besides the sample that is correctly predicted by the model, as precisely formalized in the following definition.

Definition 3 (reliability of a model) Assume that $N <$ #(members of the population).[10] The reliability of a model M is defined as

$$R(M) = \frac{\#(\text{members of the population in } M) - N}{\#(\text{members of the population}) - N},$$

where the symbol # means number of elements in the set specified in parenthesis and N is the size of the sample. ∗

[10] If $N =$ #(members of the population), we are in the extreme case that all members of the population is in the sample, in which case no reliability issue arises. At a mathematical level, condition $N <$ #(members of the population) prevents division by zero in the definition of $R(M)$. This condition in force throughout the rest of this article.

For example, if we consider the Italian population and the model $M = [0, 2\,meters]$ is used to describe the height of Italians, then we have

$$R(M) = \frac{\#(\text{Italians not taller than 2 meters}) - N}{\#(\text{Italians}) - N}.$$

Given a procedure, the model depends on the sample S, and so does the reliability of the model. For example, if we construct a model of the type $[0, \max\{height_i\}]$ and Giovanni, who is 1.92 meters tall, happens to be the tallest in the sample, then the reliability of the model is the proportion of individuals not taller than 1.92 meters; but if Marco, who is 2 meters tall, is in the sample and nobody in the sample is taller than him, then the reliability of the model rises to the proportion of individuals that are not taller than 2 meters. The reliability of a procedure based on samples of N elements is the average reliability of the models that the procedure generates when it is applied to samples of N elements. This concept is formalized in the next definition.

Definition 4 (reliability of a procedure) The reliability of a procedure P is defined as

$$R(P) = \frac{\sum_{\{S:|S|=N\}} R(P(S))}{\#(\text{subsets of the population with } N \text{ elements})}. \tag{2}$$

*

In this definition, summation is taken over all possible samples S such that $|S| = N$, that is, all possible subsets of the population with N elements.[11] The denominator equals the number of elements in the summation and it is used as a normalization factor so that $R(P)$ is the average of $R(P(S))$.

Normally, the number of terms in the summation is truly large. If e.g. $N = 100$ and the individuals are sampled from the Italian population, which is of 60 million, then the number of terms in the summation is approximately 10^{619}, a 1 followed by 619 zeros! Even if for a moment we assume that the attributes of all the Italians were known to us, then this extremely large number would nevertheless make it impractical to compute $R(P)$. On the other hand, we want to repeat once more (so as to position the nature of the problem we are studying at a level of clarity that admits no possibility of misinterpretation) that computing $R(P)$ from its definition in a real application is not just impractical, it is even impossible. Indeed, computing $R(P)$ requires to evaluate $R(P(S))$ for any S and this, in turn, entails that one has access to a complete description of the population, which is not available in practice.[12] Nonetheless, in what follows we state a proposition by which we can get around of this difficulty: in the next Proposition 1 it is shown that an abstract argument permits one to establish a fundamental link between reliability and complexity, so that reliability can be estimated from the complexity without any additional knowledge on the population. Interestingly, this result is tight in the sense that it holds with equality for certain populations and, therefore, it is not improvable.

[11] $|S|$ is the cardinality of S, i.e., the number of elements in S.

[12] Should a complete description of the population be available, then we would have nothing to learn as the attributes of the whole population would be known, and the inductive problem would not exist altogether.

Proposition 1 *Relation*

$$R(P) \geq 1 - \frac{c(P)}{N+1}$$

holds true irrespective of the population. ∗

The derivation of Proposition 1 is given in Sect. 5.2. We also note that Proposition 1 is the discrete counterpart of Theorem 1 in Calafiore and Campi (2005).

A couple of examples help clarify the result in Proposition 1.[13]

Example 3 (Enumerative induction) A classical problem in inductive reasoning takes the following form: all objects of type T observed so far have quality Q; what can I conclude about one next object of the type T that I shall observe in the future? Will it also have quality Q? For example, all pieces of bread of a certain appearance have thus far been nourishing, can I conclude that a next similar piece of bread will also be nourishing? This is known as the problem of "enumerative induction", see, e.g., Goodman (1955) and Steel (2010).

To cast enumerative induction in the setup of our procedure P, let us consider a class $\mathcal{M}$ of models that contains only two elements, $\{Q\}$ (the set formed by the sole Q) and $U = \{Q, \bar{Q}\}$, called the "universe", which contains Q and its opposite $\bar{Q}$. $\{Q\}$ is given the smallest value in the criterion of optimality (b) of Procedure P, so that $\{Q\}$ is selected if all observations have quality Q (one, e.g., makes the model that all pieces of bread of a certain appearance nourish). In the opposite, $\{Q\}$ is discarded in favor of U, i.e., both Q and $\bar{Q}$ are considered possible (this corresponds to conclude that not all pieces of bread of a certain appearance nourish if one has seen one such piece of bread that does not nourish). Observing that with no observations one chooses $\{Q\}$ and that U is chosen with only one negative observation, the conclusion can be drawn that $C(P) = 1$ in this case. With, e.g., 49 observations, applying Proposition 1 yields that this procedure has reliability at least $1 - \frac{c(P)}{N+1} = 1 - \frac{1}{50} = 98\%$. ∗

Example 2—Cont'ed Suppose that a model of the type $[0, \max\{height_i\}] \times [0, \max\{weight_i\}]$ is constructed using a sample of $N = 99$ Italians. Proposition 1 states that the reliability of the procedure is at least $1 - \frac{c(P)}{N+1} = 1 - \frac{2}{100} = 98\%$. If, instead, a Tchebyshev layer is used, the reliability is at least $1 - \frac{c(P)}{N+1} = 1 - \frac{3}{100} = 97\%$ (recall from Example 2 that $c(P) \leq 3$ in this case).

Interestingly, applying one of these two procedures to the smaller population of Luxemburg, or to the larger population of Brasil, would result in the same conclusions as for Italy. The interpretation is that informative results can be obtained for any large population, provided the question we ask is simple enough. ∗

Proposition 1 links the reliability of a procedure to its complexity, that is, to the largest possible number of dominant observations and this justifies the title of this article "Inductive knowledge under dominance".

[13] The reader may also be interested in consulting the recent monograph Campi and Garatti (2018) that contains a broad presentation of real learning problems in a context of dominance.

Referring to the formula in the proposition, function $1 - \frac{c(P)}{N+1}$ tends to 1 as N tends to infinity at a rate that is proportional to the inverse of N.[14] So, roughly (we say "roughly" because the denominator contains $N + 1$ and not just N), doubling the sample size halves the risk of incorrect prediction. The complexity of the procedure $c(P)$ sets the coefficient in the convergence rate. With a procedure that is twice as complex as another procedure we roughly need to double the sample to get the same reliability. In a learning problem, the result in Proposition 1 helps sizing the sample so that a desired level of reliability is achieved. Note also that Proposition 1 addresses issue (I) at the beginning of this Sect. 2.4.

Remark 3 Proposition 1 refers to the proportion of members in the population contained in the model, and it belongs to the branch of mathematics called *combinatorics*. In Sect. 4, this result is re-interpreted in probabilistic terms within the context of exchangeable processes under the assumption that each list of observations has the same probability to be drawn.[15] We also note that, when preparing this manuscript, we endeavored to engage the largest possible audience and, hence, we spent a substantial amount of time to leave out any continuous mathematics. On the other hand, the interested reader can consult Theorem 1 in Calafiore and Campi (2005), which provides a result in the same spirit as Proposition 1 that can be applied to generic probability spaces. ∗

3 Beyond the complexity of a procedure: reliability with model rejection

Computing $c(P)$ is not always an easy task and, when $c(P)$ is not available, Proposition 1 cannot be resorted to to determine the reliability of Procedure P. Moreover, in some cases $c(P)$ is computable but it turns out to be too large, or even equal to infinity, so that Proposition 1 is of no help to bound the reliability. For example, given a sample of points in $\mathbb{R}^2$, consider the model M given by the convex hull of the points, see Fig. 5.[16] To see that in this case $c(P) = \infty$, consider a sample S of distinct points that belong to a circle; then, certainly, all points in S are vertexes of the convex hull and removing any of them before running the procedure results in a convex hull of reduced size. Hence, $c(P, S)$ equals the cardinality of S. On the other hand, the cardinality of S can be arbitrarily large, from which it follows that $c(P) = \sup_S c(P, S) = \infty$.

[14] As we have said before the statement of Proposition 1, the evaluation of $R(P)$ provided in the proposition cannot be improved because it holds with equality for certain populations, a fact that is further commented upon in Sect. 5.2. On the other hand, it remains that the evaluation of $R(P)$ is worst-case and $R(P)$ can decay at a rate faster than the inverse of N for specific populations.

[15] This positions the result within the tradition of the *principle of indifference* (a terminology coined by John M. Keynes). While inchoate versions of this principle were already present in Blaise Pascal, Jacob Bernoulli and Gottfried W. von Leibniz, the principle of indifference was fully developed into a theoretical apparatus mainly in de Moivre (1718), and, later, in Laplace (1814).

[16] A convex set is a set where the line segment connecting any two points in the set is entirely contained in the set. Hence, a square or a disk is convex, but a horseshoe-shaped set is not. The "convex hull" of given points is the smallest convex set that contains all the points.

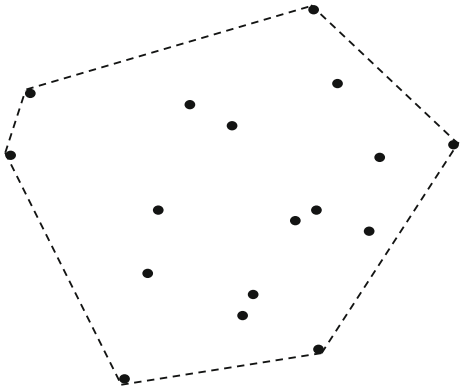

Fig. 5 The convex hull of N points

The goal of this section is to provide an alternative that can be applied when a suitable bound for $c(P)$ is not available. The idea is that one waits until after the model is constructed, and the model is accepted and used only when the complexity of the procedure for the sample at hand turns out to be not too high; otherwise, the model is judged to be too risky to use and therefore rejected. This study will lead us to give an answer to question (II) posed at the beginning of Sect. 2.4.

To be specific, suppose that after $P(S)$ is constructed, one evaluates $c(P, S)$ and, if $c(P, S)$ is below or equal to an assigned threshold c ($c < N$), the model is accepted, otherwise it is rejected. In this context, the definition of reliability needs be modified as follows (the index "c" in $R_c(P)$ refers to the threshold c):

$$R_c(P) = \frac{\sum_{\{S:|S|=N \text{ and } c(P,S)\leq c\}} R(P(S)) + \sum_{\{S:|S|=N \text{ and } c(P,S)>c\}} 1}{\#(\text{subsets of the population with } N \text{ elements})}. \qquad (3)$$

In the numerator, the reliability $R(P(S))$ is replaced by 1 (full reliability) whenever $c(P, S) > c$ since, in this case, the model is not used and the risk of incurring a wrong prediction is therefore null.

By observing that in (3) the reliability $R(P(S))$ is accounted for only when $c(P, S) \leq c$, one might expect that $R_c(P)$ cannot be lower than the reliability that hold when $c(P) = c$ (in which case the reliability $R(P(S))$ is always accounted for), which, using the bound in Proposition 1, would give $R_c(P) \geq 1 - \frac{c}{N+1}$. Somehow surprisingly, this result is incorrect as Example 4 below shows.[17] This fact is interpreted as follows:

> *if we ask a complex question with $c(P) > c$ and accept the answer only when the answer turns out to be simple enough ($c(P, S) \leq c$), then the answer is not as*

[17] A word of clarification is perhaps appropriate to dissipate any doubts regarding this claim. For a given procedure and a given population, the right-hand side of (3) does return a value that is equal to or bigger than that returned by the right-hand side of (2). The very point is that (3) applies also to populations that lie outside the domain of applicability of (2) (populations for which the complexity exceeds c). When (3) is applied to one of these populations, it can return a value larger than the upper bound given in Proposition 1 with $c(P) = c$ despite the fact that one only accepts the model when the complexity is no more than the threshold c.

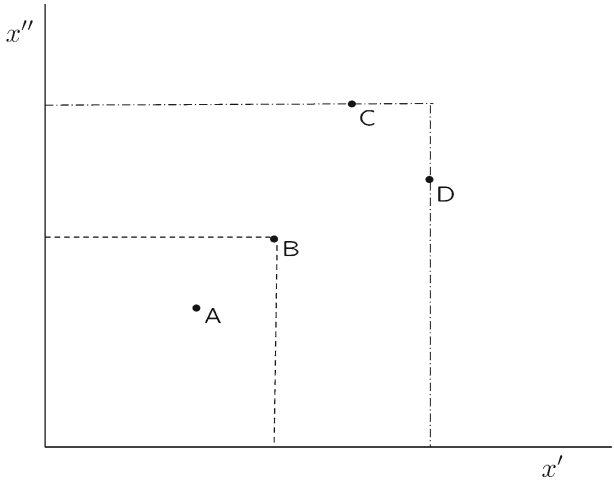

Fig. 6 Population for example showing that $R_c(P) \geq 1 - c/(N+1)$ is not a valid bound

trustworthy as when we ask a simple question ($c(P) = c$) in the first place. That is, the high complexity of a question is not compensated for by the simplicity of its answer (refer back to question (II) at the beginning of Sect. 2.4).

The margin $(1 - \frac{c}{N+1}) - R_c(P)$ is a measure of the cost one pays for having asked a complex question, even if the answer is accepted only when it turns out to be simple.

Example 4 ($R_c(P) \geq 1 - c/(N+1)$ is not a valid bound) Consider a very simple population with 4 members A, B, C, D where each member is described by 2 real attributes x' and x'' as visualized in Fig. 6. Given a sample of $N = 2$ members, suppose that M is given by the smallest rectangle with sides parallel to the coordinate axes and a vertex in the origin $(0, 0)$ that contains the sample of 2 members.
Take $c = 1$. There are 6 subsets of the population with 2 elements, which we next analyze exhaustively. If $S = \{A, B\}$, then the dashed rectangle in Fig. 6 is obtained, $c = 1$ and $R(P(S)) = 0$. If $S = \{A, C\}$ or $\{A, D\}$ or $\{B, C\}$ or $\{B, D\}$, then $c = 1$ and $R(P(S)) = 1/2$. Finally, if $S = \{C, D\}$, then the dashed-dotted rectangle is obtained and, since in this case $c = 2$, the model is not used. Hence, $R_1(P) = (0 + 1/2 + 1/2 + 1/2 + 1/2 + 1)/6 = 1/2$. However, $1 - c/(N+1) = 1 - 1/3 = 2/3 > R_1(P)$. By increasing the number of points, one can see that counterexamples to relation $R_c(P) \geq 1 - c/(N+1)$ can be found for populations of any arbitrarily large size. $*$

Hence, asking difficult questions has a price even if we listen to the answer only when the answer turns out to be simple, which is an epistemologically important fact. On the other hand, it turns out that, quantitatively, the price is modest and the following proposition delivers a lower bound for $R_c(P)$.[18]

[18] For simplicity, the proposition refers to when the cardinality of the population tends to infinity, which is referred to as the "large population" set-up. The reader unfamiliar with the notation $\binom{N}{c}$ is referred to Sect. 5.1 for an explanation. We also note that, in the formula, $\frac{1}{N-c}$ corresponds to exponentiation, that is, the binomial coefficient $\binom{N}{c}$ is raised to the fractional exponent $\frac{1}{N-c}$.

Proposition 2 *Given any procedure P, relation*

$$R_c(P) \geq \frac{1}{\binom{N}{c}^{\frac{1}{N-c}}} \cdot \frac{N-c}{N-c+1} \tag{4}$$

holds true for any "large population". *

The proof is given in Sect. 5.3.

It is useful to compare visually the lower bound given by Proposition 2 against the bound $1 - \frac{c}{N+1}$ in Proposition 1; this is provided in Fig. 7 for $c = 1, 4, 10$ and increasing values of N.

4 Discussion

4.1 Interpretation in terms of exchangeable processes

Consider an *ordered* list $(x_1, x_2, \ldots, x_{N+1})$ of $N+1$ members of the population.[19] We assume that the set of all such lists forms the space of outcomes in a probabilistic model in which each list is given the same probability. To compute this probability, observe that in a population of, say, Q members, x_1 can be any of the Q members, x_2 can be any of the Q members except x_1, and hence it is chosen from a set of $Q-1$ members, and so on till the completion of the list. Therefore, the total numbers of lists is $Q \cdot (Q-1) \cdot (Q-2) \cdots (Q-N) = Q!/(Q-N-1)!$ and the probability of each list is the inverse of this number, i.e., $p = (Q-N-1)!/Q!$ (this makes the sum of the probabilities of all lists equal to 1). In probabilistic terminology, this model is said to be *exchangeable*, a word that remarks on the fact that the probability of lists does not depend on the order in which members in the list appear.[20] In this context, the reliability $R(P)$ of procedure P (Definition 4) can be re-written as a probability as follows:

$$R(P) = \mathbb{P}\{x_{N+1} \in P(S(x_1, \ldots, x_N))\}, \tag{5}$$

where $S(x_1, \ldots, x_N)$ is the sample obtained from the list $(x_1, \ldots, x_N)$ by removing its ordering [see Sect. 5.4 for a proof of equation (5)]. The interpretation is that $R(P)$ corresponds to the probability of constructing a model from the first N members in a list of $N+1$ members and then the $(N+1)$-th member is correctly described by the model. Further, by an application of the law of large numbers (see any textbook on probability, e.g., Shiryaev (1996)) we also conclude that $R(P)$ is the long term average of successes in a repeated use of the procedure P over independent trials:

$$R(P) = \lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} 1(x_{N+1}^{(i)} \in P(S(x_1^{(i)}, \ldots, x_N^{(i)})),$$

[19] The difference between a list and a sample is that the sample is a set and does not contain a concept of ordering. Hence, two lists (a, b, c) and (b, c, a) are different, however, they correspond to the same sample.

[20] The notion of exchangeability captures, and suitably formalizes, the principle of *uniformity of nature* formulated by Hume (2008).

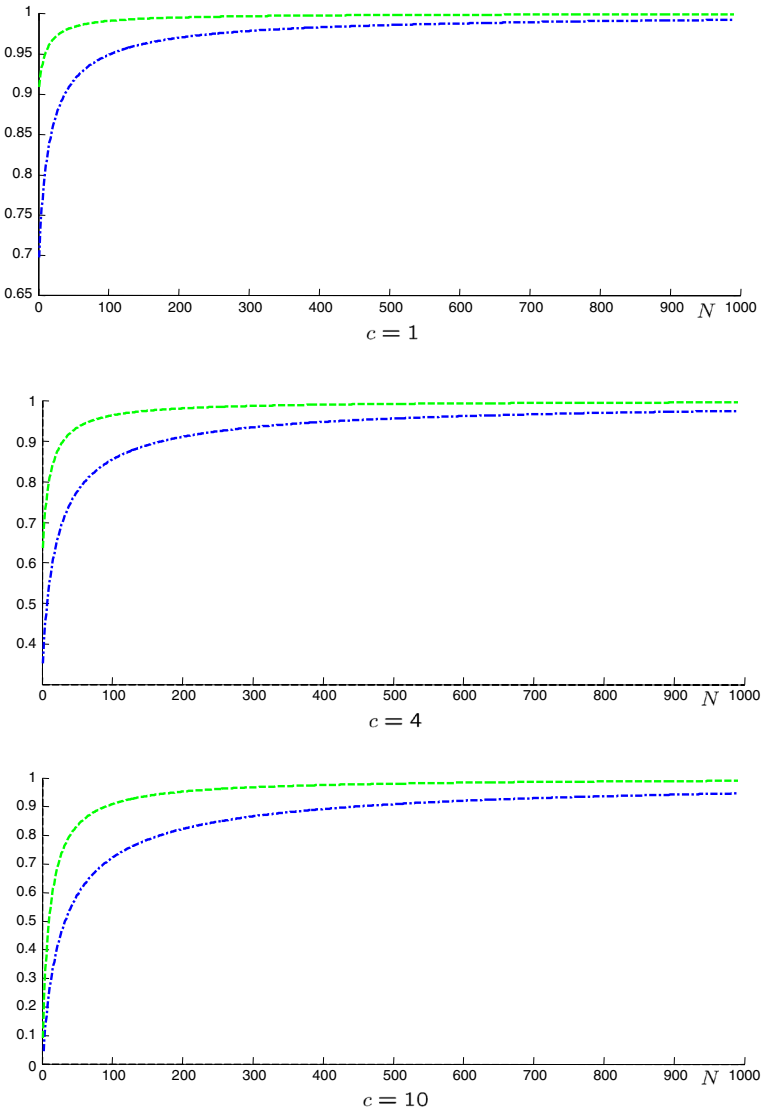

Fig. 7 The right-hand side of (4) (blue, dash-dotted graph) against $1 - \frac{c}{N+1}$ (green, dashed graph) for $c = 1, 4, 10$. The horizontal axis gives the sample size N

where superscript (i) indicates the i-th trial and convergence of the limit on the right-hand side to $R(P)$ takes place in all common probabilistic ways (e.g., with probability 1 or in mean square sense).

4.2 Interpretation in the context of Popper's theory of conjectures and refutations

In Popper's *refutation* theory, a model is *conjectured* and tested against various observations. The longer the model survives these attempts of *falsification* the more

corroborated it becomes, and it is therefore expected to survive new invalidation tests as they come along down the stream of observations, Popper (1963).

This *soft* description of Popper's refutation approach can be given a quantitative form by means of the theory of this article. In Popper's approach a model M is conjectured and provided beforehand (rather than constructed from observations). To accommodate this situation, we can consider a class $\mathcal{M}$ of models that has the conjectured model M as "base point", while $\mathcal{M}$ also contains other subsets of the population that are larger than M (that is, M is contained in all models). M is given the smallest value in the criterion of optimality (b) of Procedure P in Sect. 2.2. In this way, we accept M if all N observations are contained in M, in which case M is not falsified. Correspondingly, we can use Proposition 2 with $c = 0$ (when M survives, the complexity is zero because with no observations we choose model M) to lower bound the reliability associated with accepting the model when it is not falsified. This lower bound equals $\frac{N}{N+1}$[21] and this result can be interpreted as a quantitative statement on the level of corroboration in Popper's refutation theory. As N grows, the level of corroboration increases, converging to the value 1 when, in the limit, infinitely many observations are used to confirm the model.

Interestingly, the theory of this article also accommodates procedures that allow online updating of the model as new observations come along, so that the conjecture and refutation paradigm becomes a special case of a broader framework. To understand this, suppose that we collect observations in succession and update our model by using procedure P in Sect. 2.2 so obtaining a sequence of models $M_1, M_2, M_3, \ldots$ generated after we have obtained $N = 1, 2, 3, \ldots$ observations. If we know in advance the complexity $c(P)$ of the procedure, to all these models Proposition 1 can be applied and the reliability is lower bounded by $1 - \frac{c(P)}{N+1}$; as before, this can be interpreted as a quantitative statement on the level of corroboration and it approaches 1 as N increases, with a constant set by the value of $c(P)$. Moreover, to boost the reliability of the procedure, we can also decide to reject a model when it is too complex and Proposition 2 can be applied to this context. Following this approach, we can come up with a rigorous evaluation of the level of corroboration (regardless of whether or not the value $c(P)$ is available), so expanding the scope of Popper's analysis.[22] These results provide a justification of inductive methods by which theories are progressively updated based on observations.

5 Derivation of the results

5.1 Frequently used symbols

$1(x \in M)$ is "indicator function" of the model M, i.e. $1(x \in M)$ is 1 corresponding to members x of the population that are in M and 0 otherwise.
Given an integer L, symbol $L!$, to be read "L factorial", stands for the number obtained by multiplying all integers from L downward till 1: $L! = L \cdot (L-1) \cdots 2 \cdot 1$.

[21] Recall that $\binom{N}{0} = 1$. so that the first factor in the right-hand side of (4) is 1.

[22] In Popper's approach, a model is rejected as soon as complexity exceeds the value 0.

Symbol $\binom{Q}{N}$, to be read "Q choose N", is the number of distinct subsets of cardinality N that can be constructed from a set of cardinality Q. It turns out that $\binom{Q}{N} = \frac{Q!}{(Q-N)!N!}$. For example, the number of distinct subsets of cardinality $N = 2$ that can be constructed from a set of cardinality $Q = 3$ is 3, and this number is given by $\binom{Q}{N} = \frac{Q!}{(Q-N)!N!} = \frac{3 \cdot 2 \cdot 1}{(1)(2 \cdot 1)} = 3$.

5.2 Derivation of Proposition 1

Refer to Sect. 5.1 for the definition of symbols.

Letting Q be the size of the population, we have

$$\begin{aligned} R(P) &= \frac{\sum_{\{S:|S|=N\}} R(P(S))}{\binom{Q}{N}} \\ &= \frac{\sum_{\{S:|S|=N\}} \frac{\#(\text{members of the population in } P(S))-N}{\#(\text{members of the population})-N}}{\binom{Q}{N}} \\ &= \frac{\sum_{\{S:|S|=N\}} \frac{\sum_{\{x \in population\}} 1(x \in P(S))-N}{Q-N}}{\binom{Q}{N}} \\ &= \sum_{\{S:|S|=N\}} \frac{\sum_{\{x \in population-S\}} 1(x \in P(S))}{(Q-N) \cdot \binom{Q}{N}}, \end{aligned} \tag{6}$$

where "$population - S$" is the set of all members in the population except those in S. In the last expression, the two summations considered together range over all pairs of subsets of the population, where the first subset S has cardinality N and the second $\{x\}$ has cardinality 1 and this one is not in the first subset. The indicator function checks for the membership of x, which acts as a test member, to the model constructed from S. We can reorganize these summations by first selecting $N + 1$ members from the population in all possible ways, and then, one by one, let each of the $N + 1$ members of this subset act as a test member. In this way, (6) becomes ($G - x$ means subset G from which x is removed):

$$\frac{1}{(Q-N) \cdot \binom{Q}{N}} \sum_{\{G:|G|=N+1\}} \sum_{\{x \in G\}} 1(x \in P(G-x)). \tag{7}$$

To proceed, we consider $P(G)$, the model constructed by P when the sample is the whole G (without removing x). By Definition 1, a set of members of cardinality $c(P, G)$ from G suffices to generate model $P(G)$ using P. Let H be one such set ("one such" and not "the" because the set may not be unique). Hence, $P(H) = P(G)$. Note that if H is augmented with other members taken from G so that an enlarged subset H' is obtained, then $P(H') = P(H)$. In fact, since $P(H) = P(G)$ and $P(G)$ contains the whole subset G, $P(H)$ already contains the new members that have been added to obtain H' from H. Hence, $P(H') = P(H) = P(G)$. Consider now $G - x$; if $G - x$

contains H, then we can take $H' = G - x$ and conclude that $x \in P(G) = P(G - x)$. Therefore, $x \in P(G - x)$ holds true for at least (we say "at least" because removing a member x that is in H can still give a set $G - x$ that contains H; this may happen when G contains multiple repeats of x, i.e., the same x appear more than once in G) $N+1-c(P, G)$ choices of x. On the other hand, Definition 2 yields $c(P, G) \leq c(P)$, so that $N + 1 - c(P, G) \geq N + 1 - c(P)$. We conclude that (7) is bounded from below by the following quantity

$$\begin{aligned} &\frac{1}{(Q-N)\cdot\binom{Q}{N}} \sum_{\{G:|G|=N+1\}} (N+1-c(P)) \\ &\quad = \frac{1}{\frac{Q-N}{N+1}\cdot\binom{Q}{N}} \sum_{\{G:|G|=N+1\}} \frac{N+1-c(P)}{N+1}. \end{aligned} \tag{8}$$

Note now that

$$\frac{Q-N}{N+1}\cdot\binom{Q}{N} = \frac{Q-N}{N+1}\cdot\frac{Q!}{(Q-N)!N!} = \frac{Q!}{(Q-N-1)!(N+1)!} = \binom{Q}{N+1},$$

so that (8) becomes

$$\frac{1}{\binom{Q}{N+1}} \sum_{\{G:|G|=N+1\}} \left(1 - \frac{c(P)}{N+1}\right) = 1 - \frac{c(P)}{N+1}$$

because summation runs over all distinct subsets G of cardinality $N + 1$ constructed from a set of cardinality Q, which is equal to $\binom{Q}{N+1}$. This concludes the derivation. □

By inspecting the derivation, one can see that if, for any G, $c(P, G) = c(P)$ and set H is unique, then all inequalities in the derivation become equality so that $R(P) = 1-c(P)/(N+1)$. It is not difficult to construct examples where this happens, which shows that the result in Proposition 1 is not improvable.

5.3 Derivation of Proposition 2

Let Q be the size of the population. Given a sample S of cardinality N, let $K(S) := |P(S)| - N$, so that (compare with Definition 3)

$$R(P(S)) = \frac{K(S)}{Q-N}.$$

For any given $K \in [0, Q - N]$, we have

$$\#(S : c(P, S) \leq c \text{ and } K(S) \leq K) \leq \begin{cases} \binom{Q}{c}\binom{K+N-c}{N-c}, & \text{if } K \leq \bar{K} \\ \binom{Q}{N}, & \text{otherwise,} \end{cases} \tag{9}$$

where $\bar{K} \geq 0$ is the biggest integer such that $\binom{Q}{c}\binom{K+N-c}{N-c}$ is smaller than or equal to $\binom{Q}{N}$.[23] Proving (9) requires some attention. For an S to be counted in the left-hand side of (9), this S must first of all satisfy $c(P, S) \leq c$, that is, there exists a sub-sample of at most c members of S so that the procedure applied to this sub-sample gives $P(S)$. If this sub-sample has less than c members, then augment it with arbitrary members from S so as to get a sub-sample of cardinality c. The procedure P applied to this latter sub-sample also gives $P(S)$. Consider all possible sub-samples of cardinality c from the population. There are $\binom{Q}{c}$ such sub-samples, which we enumerate as $SS_1, \cdots, SS_{\binom{Q}{c}}$ (SS stands for "Sub-Sample"). Hence, $P(S) = P(SS_j)$ for some $j \in [1, \binom{Q}{c}]$. The second condition for an S to be counted in the left-hand side of (9) is that $K(S) \leq K$. This implies that $|P(S)| = K(S) + N \leq K + N$. As a consequence, we can bound the left-hand side of (9) as follows:

$$\begin{aligned}
&\#(S : c(P, S) \leq c \text{ and } K(S) \leq K) \\
&\quad \leq \#(S : P(S) = P(SS_j) \text{ for some } j \in \left[1, \binom{Q}{c}\right] \text{ and } |P(S)| \leq K + N) \\
&\quad \leq \sum_{j=1}^{\binom{Q}{c}} \#(S : P(S) = P(SS_j) \text{ and } |P(SS_j)| \leq K + N). \qquad (10)
\end{aligned}$$

Consider a fixed j, say $j = \bar{j}$. In order that $P(S) = P(SS_{\bar{j}})$ (first condition in (10)), the other $N-c$ members in S besides $SS_{\bar{j}}$ must be in $P(SS_{\bar{j}})$, which (second condition in (10)) is a set that has no more than $K + N - c$ elements besides $SS_{\bar{j}}$. It follows that the number of distinct ways S can be selected is at most $\binom{K+N-c}{N-c}$, i.e.,

$$\#(S : P(S) = P(SS_{\bar{j}}) \text{ and } |P(SS_{\bar{j}})| \leq K + N) \leq \binom{K + N - c}{N - c}.$$

Hence, the right-hand side of (10) is bounded by $\binom{Q}{c}\binom{K+N-c}{N-c}$. This establishes the first bound in the right-hand side of (9). The second bound $\binom{Q}{N}$ in the right-hand side of (9) is obviously true since the total number of samples S that can be formed from a population of Q members is $\binom{Q}{N}$.

Equation (9) is the fundamental building block in the derivation that follows. Write,[24]

$$\begin{aligned}
R_c(P) &= \frac{\sum_{\{S:\, c(P,S)\leq c\}} R(P(S)) + \sum_{\{S:\, c(P,S)>c\}} 1}{\binom{Q}{N}} \\
&= \frac{\sum_{\{S:\, c(P,S)\leq c\}} R(P(S)) + \sum_{\{S:\, c(P,S)>c\}} 1}{\binom{Q}{N}}
\end{aligned}$$

[23] If $\binom{Q}{c}\binom{K+N-c}{N-c} > \binom{Q}{N}$ for all $K \geq 0$, then the right-hand side of (9) is taken always to be equal to $\binom{Q}{N}$.

[24] To ease the notation, in the equation below we write the set over which summation runs as $\{S : c(P, S) \leq c\}$ instead of $\{S : |S| = N \text{ and } c(P, S) \leq c\}$, that is, the cardinality of set S is omitted; a similar shorthand applies to the the case $c(P, S) > c$.

$$+\frac{\sum_{\{S:\, c(P,S)\le c\}} 1 - \sum_{\{S:\, c(P,S)\le c\}} 1}{\binom{Q}{N}}$$
$$= 1 - \frac{\sum_{\{S:\, c(P,S)\le c\}} (1 - R(P(S)))}{\binom{Q}{N}}$$
$$= 1 - \frac{\sum_{\{S:\, c(P,S)\le c\}} \frac{Q-N-K(S)}{Q-N}}{\binom{Q}{N}}. \tag{11}$$

Note that $Q-N-K(S) = \sum_{K=0}^{Q-N-1} 1(K(S) \le K)$, so that $\sum_{\{S:\, c(P,S)\le c\}}(Q-N-K(S)) = \sum_{\{S:\, c(P,S)\le c\}} \sum_{K=0}^{Q-N-1} 1(K(S) \le K) = \sum_{K=0}^{Q-N-1} \sum_{\{S:\, c(P,S)\le c\}} 1(K(S) \le K) = \sum_{K=0}^{Q-N-1} \#(S : c(P, S) \le c \text{ and } K(S) \le K)$, which, using (9), is bounded by

$$\sum_{K=0}^{\bar{K}} \binom{Q}{c}\binom{K+N-c}{N-c} + \sum_{K=\bar{K}+1}^{Q-N-1} \binom{Q}{N}. \tag{12}$$

Using this expression in (11) yields:

$$R_c(P) \ge 1 - \frac{1}{Q-N}\frac{\binom{Q}{c}}{\binom{Q}{N}} \sum_{K=0}^{\bar{K}} \binom{K+N-c}{N-c} - \frac{Q-N-1-\bar{K}}{Q-N}$$
$$= \frac{1+\bar{K}}{Q-N} - \frac{1}{Q-N}\frac{\binom{Q}{c}}{\binom{Q}{N}} \sum_{K=0}^{\bar{K}} \binom{K+N-c}{N-c}. \tag{13}$$

Term $\sum_{K=0}^{\bar{K}} \binom{K+N-c}{N-c}$ can be further bounded as follows (the first equality is the hockey-stick identity)

$$\sum_{K=0}^{\bar{K}} \binom{K+N-c}{N-c} = \binom{\bar{K}+1+N-c}{N-c+1}$$
$$\le \frac{(\bar{K}+1+N-c)^{N-c+1}}{(N-c)!(N-c+1)}, \tag{14}$$

which, used in (13), gives

$$R_c(P) \ge \frac{1+\bar{K}}{Q-N} - \frac{1}{Q-N}\frac{\binom{Q}{c}}{\binom{Q}{N}} \frac{(\bar{K}+1+N-c)^{N-c+1}}{(N-c)!(N-c+1)}$$
$$= \frac{1+\bar{K}}{Q-N} - \frac{1}{Q-N}\frac{\frac{Q!}{(Q-c)!c!}}{\frac{Q!}{(Q-N)!N!}} \frac{(\bar{K}+1+N-c)^{N-c+1}}{(N-c)!(N-c+1)}$$
$$= \frac{1+\bar{K}}{Q-N} - \frac{1}{Q-N}\frac{(Q-N)!}{(Q-c)!}\binom{N}{c}\frac{(\bar{K}+1+N-c)^{N-c+1}}{N-c+1}. \tag{15}$$

To proceed, we evaluate $\bar{K}$ and show the validity of relation

$$\bar{K} = \frac{1}{\binom{N}{c}^{\frac{1}{N-c}}} Q + o(Q), \tag{16}$$

where $o(Q)$ stands for a quantity that grows with Q less than linearly or, in formulae, $\lim_{Q\to\infty} o(Q)/Q = 0$. To show (16), recall that $\bar{K} \geq 0$ is the biggest integer such that

$$\binom{Q}{c}\binom{K+N-c}{N-c} = \frac{Q!}{(Q-c)!c!}\frac{(K+N-c)(K+N-c-1)\cdots(K+1)}{(N-c)!}$$

is smaller than or equal to

$$\binom{Q}{N} = \frac{Q!}{(Q-N)!N!},$$

which, after reorganizing the terms, becomes

$$\begin{aligned}\binom{N}{c}\ &(K+N-c)(K+N-c-1)\cdots(K+1)\\ &\leq (Q-c)(Q-c-1)\cdots(Q-N+1).\end{aligned}$$

Dividing both sides of this inequality by Q^{N-c}, with the position $\alpha := K/Q$ one obtains

$$\begin{aligned}\binom{N}{c}&\left(\alpha+\frac{N-c}{Q}\right)\left(\alpha+\frac{N-c-1}{Q}\right)\cdots\left(\alpha+\frac{1}{Q}\right)\\ &\leq \left(1-\frac{c}{Q}\right)\left(1-\frac{c+1}{Q}\right)\cdots\left(1-\frac{N-1}{Q}\right).\end{aligned} \tag{17}$$

α takes on rational value K/Q. As $Q\to\infty$, α can approach any real in $[0, 1]$ so that in the limit when $Q\to\infty$ the largest α satisfying (17) is the solution of the polynomial equation obtained from (17) by setting to zero the terms that vanish as $Q\to\infty$, this polynomial is

$$\binom{N}{c}\alpha^{N-c} = 1.$$

Hence, with the notation $\bar{K}(Q)$, which emphasizes the dependence of $\bar{K}$ on Q, we have $\bar{K}(Q)/Q =: \bar{\alpha}(Q) \to 1/\binom{N}{c}^{\frac{1}{N-c}}$ as $Q\to\infty$, or, equivalently, $\bar{K}(Q) = Q/\binom{N}{c}^{\frac{1}{N-c}} + o(Q)$, which is (16).

Now, to obtain the result in Proposition 2 substitute (16) in equation (15) to obtain

$$R_c(P) \geq \frac{1+\frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}Q+o(Q)}{Q-N} - \frac{1}{Q-N}\frac{(Q-N)!}{(Q-c)!}\binom{N}{c}\frac{\left(\frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}Q+o(Q)+1+N-c\right)^{N-c+1}}{N-c+1}. \tag{18}$$

As $Q \to \infty$, the first term in the right-hand side of (18) tends to $1/\binom{N}{c}^{\frac{1}{N-c}}$; the second term can instead be handled as follows:

$$\begin{aligned}
&\frac{1}{Q-N}\frac{(Q-N)!}{(Q-c)!}\binom{N}{c}\frac{\left(\frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}Q+o(Q)+1+N-c\right)^{N-c+1}}{N-c+1} \\
&= \frac{1}{Q-N}\frac{1}{(Q-c)\cdots(Q-N+1)}\binom{N}{c}\frac{\left(\frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}Q+o(Q)\right)^{N-c+1}}{N-c+1} \\
&\quad \text{(term } 1+N-c \text{ has been incorporated in } o(Q)) \\
&= \frac{Q}{Q-N}\frac{Q^{N-c}}{(Q-c)\cdots(Q-N+1)}\binom{N}{c}\frac{\left(\frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}Q+o(Q)\right)^{N-c+1}}{Q^{N-c+1}}}{N-c+1},
\end{aligned}$$

where the last expression has been obtained by multiplying the numerator and the denominator by Q^{N-c+1}. As $Q \to \infty$, this latter expression tends to $\binom{N}{c}\frac{1}{\binom{N}{c}^{\frac{N-c+1}{N-c}}}\frac{1}{N-c+1} = \frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}\frac{1}{N-c+1}$. Using the results obtained for the first term and the second term in (18), we conclude that, as $Q \to \infty$ (large population), $R_c(P)$ satisfies the relation

$$R_c(P) \geq \frac{1}{\binom{N}{c}^{\frac{1}{N-c}}}\left(1-\frac{1}{N-c+1}\right),$$

and this concludes the derivation of the proposition. □

5.4 Derivation of equation (5)

Note first that there are $N!$ lists of N members that give the same sample. In fact, a permutation of the members of the list does not change the sample and the total

number of permutations is $N!$: one first chooses a member from a collection of N members to be placed in first position, then chooses a member to be placed in second position from the collection of the remaining $N-1$ members and so on corresponding to an overall number of choices given by $N \cdot (N-1) \cdot (N-2) \cdots 1 = N!$. Hence, the expression $\sum_{\{S:|S|=N\}} R(P(S))$ appearing in Definition 4 of $R(P)$ can also be written as $\frac{\sum_{\{\text{list}:|\text{list}|=N\}} R(P(S(\text{list})))}{N!}$, where $S(\text{list})$ is the sample generated from a list by removing its ordering. We therefore have:

$$
\begin{aligned}
R(P) &= \frac{\sum_{\{S:|S|=N\}} R(P(S))}{\#(\text{subsets of the population with } N \text{ elements})} \\
&= \frac{\sum_{\{\text{list}:|\text{list}|=N\}} R(P(S(\text{list})))}{N! \cdot \#(\text{subsets of the population with } N \text{ elements})} \\
&= \frac{\sum_{\{\text{list}:|\text{list}|=N\}} \frac{\#(\text{members of the population in } P(S(\text{list})))-N}{Q-N}}{N! \cdot \binom{Q}{N}} \\
&\qquad\qquad \text{[use Definition 3]} \\
&= \frac{(Q-N-1)!}{Q!} \sum_{\{\text{list}:|\text{list}|=N\}} [\#(\text{members of the population in } P(S(\text{list}))) - N].
\end{aligned}
\tag{19}
$$

Note now that, given any list of N members, quantity [#(members of the population in $P(S(\text{list}))) - N$] can be evaluated by referring to all lists of $N+1$ members that are obtained from the original list of N members by augmenting this list with one more arbitrary member that was not in the original list and then counting in how many cases this added member is in the model generated by the original list. This gives:

$$
\begin{aligned}
&\sum_{\{\text{list}:|\text{list}|=N\}} [\#(\text{members of the population in } P(S(\text{list}))) - N] \\
&\quad = \sum_{\{\text{list}:|\text{list}|=N+1\}} 1(x_{N+1} \in P(S(x_1, \ldots, x_N))),
\end{aligned}
$$

which, used in (19), finally gives

$$
R(P) = p \cdot \sum_{\{\text{list}:|\text{list}|=N+1\}} 1(x_{N+1} \in P(S(x_1, \ldots, x_N))),
$$

where we have also used the fact that $(Q-N-1)!/Q!$ is the probability p of each list of $N+1$ members. The derivation of the result is completed by noting that the expression on the right-hand side is the probability of the event $\{x_{N+1} \in P(S(x_1, \ldots, x_N))\}$.

Acknowledgements The author would like to thank Dr. Sean Kenny for providing suggestions on how to improve the presentation of this work. The author also gratefully acknowledges the valuable and constructive comments made by anonymous referees.

Funding No funds, grants, or other support was received.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
Bailer-Jones, D. (2009). *Scientific models in philosophy of science*. University of Pittsburgh Press.
Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression*. Wiley.
Calafiore, G., & Campi, M. (2005). Uncertain convex programs: Randomized solutions and confidence levels. *Mathematical Programming, 102*(1), 25–46.
Campi, M., & Garatti, S. (2018). Introduction to the scenario approach. In *MOS-SIAM series on optimization*.
Campi, M., S, S. G., & Ramponi, F. (2018). A general scenario theory for nonconvex optimization and decision making. *IEEE Transactions on Automatic Control, 63*, 4067–4078.
Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford University Press.
Chaitin, G. (1966). On the length of programs for computing binary sequences. *Journal of the ACM, 13*, 547–569.
Contessa, G. (2007). Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science, 74*, 48–68.
Da Costa, N., & French, S. (2000). Models, theories, and structures: Thirty years on. *Philosophy of Science, 67*, 116–127.
de Finetti, B. (1989). Probabilism: A critical essay on the theory of probability and on the value of science. *Erkenntnis* (translation of "Probabilismo Saggio critico sulla teoria delle probabilita e sul valore della scienza", Biblioteca di Filosofia, Napoli, 1931) *31*, 169–223.
de Moivre, A. (1718). The doctrine of chances: Or, a method of calculating the probability of events in play. W. Pearson (reprinted 1967, New York, NY: Chelsea).
Forster, M., & Sober, E. (1994). How to tell when simple, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1–35.
Forster, M., & Sober, E. (2011). AIC scores as evidence: A Bayesian interpretation. In M. Forster & P. Bandyopadhyay (Eds.), *Philosophy of statistics (Handbook of the philosophy of science)* (Vol. 7, pp. 535–549). Elsevier.
Frigg, R., & Hartmann, S. (2012). Models in science. In Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*. Fall 2012
Goodman, N. (1955). *Fact, fiction, & forecast*. Harvard University Press.
Harris, T. (2003). Data models and the acquisition and manipulation of data. *Philosophy of Science, 70*, 1508–1517.
Harter, H. (1982). Minimax method. *Encyclopedia of statistical sciences* (Vol. 4, pp. 514–516). Wiley.
Hughes, R. (1997). Models and representation. *Philosophy of Science, 64*, 325–336.
Hume, D. (2008). *An enquiry concerning human understanding*. Oxford World Classics (originally published in 1748).
Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission (USSR), 1*, 4–7.
Kolmogorov, A. (1968). Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory, 14*, 662–664.
Laplace, P. (1814). Essai philosophique des probabilitès. (translated version Philosophical Essay of Probabilities, Springer, 1999).
Laymon, R. (1982). Scientific realism and the hierarchical counterfactual path from data to theory. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association* (pp. 107–121).
Magnani, L., & Nersessian, N. (Eds.). (2002). *Model-based reasoning: Science, technology, values.* Kluwer.
Magnani, L., Nersessian, N., & Thagard, P. (Eds.). (1999). *Model-based reasoning in scientific discovery.* Kluwer.

Maki, U. (1994). Isolation, idealization and truth in economics. In Hamminga, B., & Marchi, N.D. (eds) *Idealization VI: Idealization in economics. Poznan studies in the philosophy of the sciences and the humanities* (Vol. 38, pp. 147–168). Rodopi, Amsterdam.
Mayo, D. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.
McAllister, J. (1997). Phenomena and patterns in data sets. *Erkenntnis, 47*, 217–228.
Morgan, M., & Morrison, M. (1999). Models as mediating instruments. In M. Morgan & M. Morrison (Eds.), *Models as mediators. Perspectives on natural and social science* (pp. 10–37). Cambridge University Press.
Popper, K. (1963). Conjectures and refutations: The growth of scientific knowledge. Routledge & Kegan Paul.
Shiryaev, A. (1996). *Probability*. Springer.
Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.
Sober, E. (2015). *Ockham's razors*. Cambridge University Press.
Steel, D. (2010). What if the principle of induction is normative? Formal learning theory and Hume's problem. *International Studies in the Philosophy of Science, 24*, 171–185.
Suppes, P. (1960). A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese, 12*, 287–301.
Suppes, P. (1962). Models of data. In: Nagel, E.P.S., & Tarski, A. (Eds.), Methodology and philosophy of science: Proceedings of the 1960 international congress. Stanford University Press (pp. 252–261).
Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*, 449–508.
Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences), 153*, 12–18. in Japanese.
van Fraassen, B. (1980). *The scientific image*. Oxford University Press.
Woodwart, J. (1989). Data and phenomena. *Synthese, 79*, 393–472.