Algorithmic Information Theory – A Consistent Approach to Randomness and Hypothesis Testing: An Overview of Ideas

Vladik Kreinovich, Olga Kosheleva, and Chon Van Le

Abstract The traditional statistical approach to hypothesis testing is based on the idea that events with very small probability cannot happen. The problem is that the usual naive formalization of this approach is, in general, inconsistent. This inconsistency, in its turn, often leads to irreproducibility and inadequacy of the results. In this paper, we show a new understanding of randomness – known as Algorithmic Informal Theory – can help resolve these problems. By the way, this new approach was pioneered, in the mid-1960s, by Andrei Kolmogorov, the same mathematician who, in the mid 1930s, pioneered a formalization of probability theory.

1 What this paper is about

In the last decades, many publications have emphasized that applied statistics is facing several related crises. From the theoretical viewpoint, many of its traditionally used methods face a *foundational* crisis: their conclusions are sometimes inconsistent. From the practical viewpoint, applied statistics faces two crises:

- an *adequacy* crisis – its conclusions are sometimes inconsistent with the practice of the corresponding applied discipline, and
- a *reproducibility* crisis – its conclusions are sometimes not supported by practice.

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: olgak@utep.edu

Chon Van Le
International University, Vietnam National University — Ho Chi Minh City, Vietnam
e-mail: lvchon@hcmiu.edu.vn

The foundational problems have been known for the longest time. In the 1960s none else but Andrei Kolmogorov – a mathematician who, in the 1930s, pioneered mathematical foundations for probability theory – noticed the related problems and pioneered what is now called Algorithmic Information Theory (also known as Kolmogorov complexity), a supplement to the traditional probability theory.

In this paper, we start with a reminder, in Section 2, of how the statistics has been traditionally applied. Then, in Section 3, we briefly explain the three crises. In Section 4, we explain how the need to resolve these crises has led Kolmogorov and others to the main ideas of Algorithmic Information Theory, ideas that can be helpful in solving the crises. In Section 5, we explain the remaining challenges related to potential applications of the new approach.

2 How statistics has been traditionally applied

How science works: a brief reminder. One of the main objectives of science is to be able to predict the future state of the systems – based on the current and past observations and measurements.

To be able to do it, we need to understand how the future state of a system depends on its previous states. Such dependencies are known as *laws of nature*. A classical example is Newton's mechanics, where Newton's equations enable us to predict the future state of mechanical systems – such as the Solar system.

So, to be able to make predictions, researchers discover – and test – laws of nature, and then use the discovered laws to make predictions.

How are laws discovered? *Discovering* a dependence between physical quantities based on observations is still a creative process – this is what Nobel prizes are given for.

Need for testing. Once a hypothetical law is formulated, we need to *test* it, to check whether it is consistent with all the observations and measurement results.

How can we test: deterministic case. In application areas like mechanics, where the systems are deterministic, and where, in principle, exact observations are possible – so that the measurement errors can be safely ignored – testing is easy: we expect the future observations to coincide (within a small measurement error) with the theory's prediction.

If a planet is observed exactly at the location that was predicted, this means that the theory is confirmed. On the other hand, if the planet's actual location is different from what we predicted, the theory needs to be modified.

Enter randomness. Deterministic testing helps when the following two conditions are satisfied:

- first, the future state of the system is uniquely determined by its current state – and, moreover, uniquely determined by the values of the observed quantities, and
- second, all the measurements are observations are accurate.

In many real-life systems, at least one of these conditions is not satisfied:

- first, the future state of the system is often also depending on factors that we do not observe; such situations are typical in economics, where the future economic state depends also on politics, on new inventions, on weather, etc.;
- second, measurements and observations are often very approximate.

In such cases, we cannot predict the exact values of the future observations and measurement results – at best, we can predict the *probabilities* of different future measurement results.

How can we test: the general case, when only probabilistic predictions are possible. In such a general case, we cannot simply compare the measurement results and the hypothesis's predictions, we need to be able to test the corresponding hypothesis in the case of randomness.

How traditionally hypotheses have been (and sometimes are) tested. Traditional methods for testing hypotheses – methods which are often summarized under the "p-value" rubric – usually select some small probability value p_0. Then:

- If we have an observed event E whose probability $p(E\,|\,H)$ under the given hypothesis H is smaller than or equal to p_0, we conclude that the hypothesis is rejected.
- On the other hand, if $p(E\,|\,H) > p_0$, then we conclude that the hypothesis is not rejected.

Example. This event E is usually an inequality between come combination of observed values and a fixed number.

For example, suppose that we have a class of objects with known sample mean m and known standard deviation s, and we need to check whether a new object with the value x belongs to this class.

Let us be even more specific. Suppose that we have a population of people who are mostly reasonably healthy – with different values of body temperature. Based on this data, we want to come up with a method that will use the current person's body temperature to decide whether he/she is healthy or not.

In such a situation, it is reasonable to compare the ratio

$$z \stackrel{\text{def}}{=} \frac{x-m}{s},$$

which is known as a *z-score*, with some threshold z_0. For example, if $|z| > 3$, i.e., if z is outside 3-sigma interval $[m-3s, m+3s]$, then we conclude that the person's body temperature is abnormal.

Comment. This is, by the way, how most "healthy" ranges are constructed in medicine.

Main idea behind the traditional approach. One can see that the main idea behind the traditional approach to testing can be described very simply: events with very small probability cannot happen.

At first glance, this idea is in perfect agreement with common sense. This idea seems to be in perfect accordance with the common sense.

For example, in a country's lottery in which millions of people participate, someone wins the main prize – this is understandable. But if the same person wins the main prize two years in a row, people will be naturally suspicious of fraud – because the probability of such an event is very small.

If you flip a coin and it falls head 100 times in a row, you will naturally conclude that this coin is biased – because the probability 2^{-100} of this event is too small.

If we see a man rising in the air by himself, we suspect that there is some invisible apparatus that helps him. In principle, it is possible that all randomly moving molecules in a person's body start moving in the same vertical direction – but the probability of this is very small, so we do not consider this option as possible.

On the other hand. In the following sections, we will see, however, that:

- while on the qualitative level, this idea agrees with common sense,
- a simplified implementation of this idea in the p-value approach – that all events with probability $p \le p_0$ are not possible – has many problems.

Comment. The fact that the traditional p-value approach has many problems is well known; see, e.g., [1, 12] – but is still worth repeating, because many researchers still continue to use this approach. We will repeat these problems in this paper as well. However, our main focus is *not* on these problems, but on Kolmogorov's approach to *resolve* these problems.

3 The three crises

3.1 What we do in this section

As we mentioned in the previous section, the traditional ("p-value") approach to hypothesis testing is based on the seemingly natural idea that events with small probability – probability smaller than some very small threshold p_0 – cannot happen.

In this section, we show that this idea has many problems: it is inconsistent, it is not adequate, and it leads to irreproducible results.

3.2 Foundational crisis

What this crisis is about. The foundational crisis is the easiest to explain.

Suppose that we flip a fair coin N times in a row. No matter what is the resulting sequence of heads (H) and tails (T), the probability of each such sequence is the same: 2^{-N}.

So, no matter how small the threshold p_0, for a sufficient large N, we have $2^{-N} < p_0$. So, none of the 2^N events is possible. So, the above hypothesis would imply that none of the N-long sequences of coin flips is possible at all – but we can flip a coin as many times as we want and thus, get a quite possible sequence.

This means that the above assumption – that events with probability $\le p_0$ are not possible – is inconsistent. By simply flipping a coin sufficiently many times, we can prove that this assumption cannot be applied to all possible events.

This crisis cannot be resolved by lowering p_0. Traditional approach to hypothesis testing uses $p_0 = 0.05$. At first glance, one may think that the problem is caused by the fact that this threshold is too high: we can gave many events with lower probability. For example, physicists do not use $p_0 = 0.05$. Instead, they use the 5-sigma criterion of rejecting a hypothesis – that values outside the interval

$$[m - 5\sigma, m + 5\sigma]$$

are not possible. This criterion corresponds to $p_0 \approx 6 \cdot 10^{-7}$.

This lowering helps, but, as we have argued in the previous subsection, it does not solve the foundational problem – the problem appears no matter how small is p_0.

3.3 Reproducibility crisis: with or even without p-hacking

How science works: a reminder. To explain this crisis, let us recall how researchers work. A researcher trying to understand a phenomenon deduces a hypothesis – based on some data. Then, he/she tests whether this hypothesis fits all available observations (and/or new experiments). If it does fit, he/she publishes a paper.

If the hypothesis does not fit some data, a natural idea is to modify the hypothesis based on the new observation – and to test whether the modified hypothesis will work better. This is how science operates. First, Newton came up with his equations of motion. Then, it turned out that his equations are not valid for large velocities and for small sizes – and so relativity theory and quantum physics were born.

With p-values, this leads to the reproducibility crisis. Somewhat surprisingly, when we combine this very natural approach to doing science with the p-value techniques, we get a problem. Indeed, the usual standard for hypothesis rejection is using the threshold value $p_0 = 0.05$. This value means that for a false hypothesis, there is a 5% probability that it will be not rejected.

This may be tolerable, but what if we have 20 false hypotheses tested one after another? Then, the probability that one of these hypotheses will be not rejected is equal to 1 minus the probability that all of them will be rejected, i.e., to

$$p = 1 - \left(1 - \left(\frac{1}{20}\right)\right)^{20}.$$

It is known that when n tends to infinity, then

$$\left(1-\frac{1}{n}\right)^n$$

tends to e^{-1}, where $e = 2.7129828\ldots$. Thus, we have $p \approx 1 - e^{-1} \approx 62\%$.

So, by following the usual way of doing science, in more than half of the cases, we accept a false hypothesis. How will we know that the accepted hypothesis is false? Easily: it does not fit the new data, i.e., the result will not be reproducible. This is exactly what happened when researchers tried to check how many published results in biology, medicine, and other areas are reproducible: it turns out that more than half of them are not reproducible; see, e.g., [4].

Is this just an ethical problem of p-hacking? In some cases, false hypotheses come from simply trying all possible hypotheses – in the hope that one of them sticks. As we have just argued, this will eventually lead to a hypothesis which is not rejected – even if actually all tests hypotheses are false. This practice – called *p-hacking* – has indeed been used by some researchers.

Now that we know that this practice leads to an acceptance of a false hypothesis, this practice is unethical – especially in fields like medicine, where the acceptance of a false hypothesis can lead to wrong treatments and even death. From this viewpoint, this seems like an ethical problem – do not do p-hacking, and everything will be OK.

Unfortunately, the same problem occurs for perfectly ethical researchers as well. Indeed, in line with the above argument, for a false hypothesis to be accepted, there is no need for the same researcher to test 20 hypotheses. For example, in medicine, where problems are practically important, there are usually several researchers trying to understand a certain phenomenon.

If 20 different researchers propose 20 different hypotheses, then, even if actually all of them are false, there is more than 60% probability that one of these false hypotheses will be accepted – and none of these 20 researchers did anything unethical, they all operated in good faith.

Can Bayesian approach help? In general, it is a good idea to take into account not only current observations, but also prior knowledge – that can be described by prior probabilities. This is very important in all areas, e.g., in medicine, where a skilled doctor takes into account not only the values of the patient's blood pressure, etc., but also the doctor's multi-year experience.

From this viewpoint, the Bayesian approach – that takes this prior knowledge into account – often leads to better practical results. In this case, we reject a hypothesis only if the Bayesian-motivated probability p of an event is smaller than a given threshold p_0. This indeed usually leads to better practical results. However, it does not solve the fundamental problem of p-hacking: indeed, with the same threshold p_0, we have the exact same probability of accepting a false hypothesis.

Can using physics-based approach help? In economics and finance, a hypothesis is often formulated in terms of a probability distribution – e.g., in terms of the probability distribution of the difference between the actual and predicted values of an economic quantity.

In contrast, in physics, a hypothesis is usually rejected or accepted based on how well it predicts the future values. It is a completely different approach, but, from the foundational viewpoint, this approach has the same problems: we still need to decide when there is a fit between the prediction and observations, and for that, we need to use some hypothesis testing technique.

3.4 Crisis of adequacy

The last crisis can be best illustrated on the above example of a long sequence of coin flips – or on an example of a person floating in air because of the random molecules randomly moving in the same direction.

From the viewpoint of mathematical statistics, it is quite possible that we flip a fair coin 100 times in a row, and we get heads every time – the only caveat is that since such events are very rare, we need to wait for a very long time for this to happen. Similarly, according to probability theory, it is possible for a person to suddenly start floating in air because of the random molecular motion – it will just take a lot of time for us to observe this.

However, if you ask a physicist, he/she will say that such rare events are simply *not* possible; see, e.g., [2, 11]. From the physicists' viewpoint:

- the theoretical probability – that due to random motion a mixture of water and oil will spontaneously separate – may be positive,
- but in practice, such a separation is simply *not* possible.

We may ignore the opinion of the physicists by claiming that they are simply ignorant of probability theory. However, let us be cautious about such an approach. After all, physicists have had many successes in explaining the world, successes that have led to many useful gadgets. So, it is probably a good idea to take their intuition seriously – even if it seems to contradict our mathematics.

3.5 So what can we do?

As we have mentioned, in the 1960s, Kolmogorov and others did take the physicists' intuition seriously, and they came up with a consistent approach to the notions of randomness and to hypothesis testing, a promising approach that we will overview in the next section.

Comment. In the following section, we will describe the main ideas behind this approach; a detailed description can be found in [10].

4 Main ideas behind Algorithmic Information Theory

4.1 What is a random sequence?

What do we mean when we say that a sequence of heads (H) and tails (T) – obtained by flipping a fair coin – is random?

From the viewpoint of the traditional probability theory, this very question is a taboo. Many textbooks continue to claim that there is no such thing as a random number or a random sequence, all we have is a probability distribution over the set of all the numbers – or over the set of all the sequences.

However, from the commonsense viewpoint, the actual sequence obtained by slipping a coin (or, better yet, by running a quantum process for which the probability of an outcome is 1/2) is random, while periodic sequences TT... or THTHTH... are clearly not random. How can we formalize this difference?

To answer this question, let us recall why we think that the above two sequences TT... and THTHTH... are not random. The sequence TT... is not random because in a truly random sequence, the frequency of Ts should be 1/2. Why? Because for almost all sequences, the frequency is 1/2. Similarly, the sequence THTH... is not random because the frequency of a subsequence TT should be 1/4.

Frequencies are not the only criterion. We can have a sequence in which all frequencies are correct, but if the deviations of the frequencies from their mean values do not follow the normal distribution – as is true for almost all sequences – this sequence is not random.

We can summarize this by saying, informally, that an infinite sequence is random if it satisfies all laws of probability, i.e., all statements that hold, for a given probability distribution, with probability 1. Equivalently, this means that a sequence is random if it does not belong to any *definable* set of measure 0.

By definable, we mean that there is a statement (e.g., in set theory or in any other formalization of mathematics) that uniquely determines this set. For example, the set of all the sequences with a given frequency of Ts is definable, the set of all the sequence with a given frequency of TTs is definable, etc. However the set consisting a single element – the actual sequence of coin flipping results – is not definable.

In any formal language, there are countably many possible statements, so there are countably many definable sets. It known that the union of countably many sets of measure 0 also have measure 0. Thus, the union of all definable sets of measure 0 also has measure 0. So, as expected, almost all sequences are random.

Let us describe this in precise terms.

Definition 1. *Let X be a definable set, and let p be a definable probability measure on this set.*

- *By a* probability law *we mean a definable statement whose probability is 1.*
- *An element $x \in X$ us called* random *if it satisfies all probability laws.*

Comments.

- This is equivalent to saying that an element $x \in X$ random if and only if x does not belong to any definable subset of measure 0.
- Depending on what formal language we use, we may get slightly different definitions of randomness. For example, if we limit ourselves to statements formalized in arithmetic or in mathematical analysis, then we may have a sequence as random in the sense of this definition. However, this sequence may not be random for other formal languages: e.g., it may be an element of a set of measure 0 that can only be defined in the language of set theory.

Proposition 1. *For every definable probability measure on a definable set, almost all elements are random.*

Comment. This is what we have proved earlier in this section.

4.2 What is a finite random sequence?

The previous subsection defined when an infinite sequence is final, but intuitively, we also apply this term to final sequences. The actual sequence of 100 coin flips is random, but the sequence THTH... of length 100 is clearly perceived as not random – but why?

Kolmogorov came up with a natural answer to this question: a sequence THTH... can be described by a very short statement – that it is TH repeated 50 times. In contrast, we do not expect the actual sequence of coin flip results to be describable by any short statement – it is random because it is kind of lawless.

This idea led to the following natural definitions. These definitions use the fact that in a computer, everything – any statement, any program – is represented as a binary sequence, i.e., a sequence of 0s and 1s.

Definition 2. *Let a formal language L be given. By a* Kolmogorov complexity $K(x)$ *of a finite string x we mean the shortest length of a definition of x in the language L.*

Historical comments. This idea was also independently discovered by two other researchers: Ray Solomonoff and Gregory Chaitin.

This definition was first formulated for the language of algorithms and computer programs. In this case, $K(x)$ is simply the shortest length of a program that generates the string x. Because of this, the whole researcher area was called *Algorithmic Information Theory* – and this name is still in use, although the definitions have been extended to non-algorithmic situations as well.

Discussion. The above definition depends on which formal language we use. For example, we can use one of the programming languages – as in the original definition by Kolmogorov and others. Alternatively, we can use the language of set theory, etc.

If we have a description in a formal language L, then we can have a description in another language L' with the same expressive ability if we add, to the description in L, the definition of L in terms of L'. Thus, the shortest definition $K_{L'}(x)$ of x in L' cannot exceed the sum $K_L(x)+C$, where C is the length of the translation of L into L'. Similarly, we have $K_L(x) \le K'_L(x)+C'$ for some constant C'.

So, Kolmogorov complexity is defined modulo an additive constant: for equivalent formal languages, for some constant C, we have $|K_L(x)-K_{L'}(x)| \le c$ for all x, where we denoted $c \stackrel{\text{def}}{=} \max(C,C')$.

Reminder. A binary sequence x is random if the shortest description of this sequence consists of simply listing all its values, i.e., has the length close to the length $\text{len}(x)$ of this sequence.

In a long random sequence, we may have a few cases when we have, e.g., a repetition – because of which we can shorten the description of this random sequence. For example, we may have several 1s in a row, in which case it may be easier to say that all the bits in this subsequence are 1s instead of listing all these bits. Indeed, instead of listing b equal bits, we can simply describe the number b – and we need only $\log_2(b) \ll b$ bits to describe gthe number b.

It is reasonable to expect that the number of such repetitions is proportional to the length of the random sequence, with some small proportionality coefficient ε. So, it makes sense that assume that when a finite sequence x is random, we must have $K(x) \ge (1-\varepsilon)\cdot \text{len}(x)$.

Let us take into account that the Kolmogorov complexity is defined modulo an additive constant. So, if for one of its definitions, we have the above inequality, then for other definitions of $K(x)$, we have $K(x) \ge (1-\varepsilon)\cdot \text{len}(x) - c$ for some constant c. So, it makes sense to come up with the following definition.

Definition 3. *Let c and ε be positive ina positive integer. We say that a binary sequence x is* (c,ε)-random *if* $K(x) \ge (1-\varepsilon)\cdot \text{len}(x) - c$.

4.3 How is this related to hypothesis testing: discussion

Based on the above definition, how can we formalize the idea that events with very small probability cannot happen? According to the above definition of randomness, if we assume that a sequence is random, this means that some binary sequences of length n are not possible. Namely, if $K(x) < (1-\varepsilon)\cdot n - c$, then the sequence x is not possible.

Let us describe this inequality in terms of the probabilities. The probability of each binary sequence of length n is $p = 2^{-n}$. The above impossibility-inducing inequality $K(x) < (1-\varepsilon)\cdot n - c$ is equivalent to

$$n > \frac{K(x)}{1-\varepsilon} + c',$$

where we denoted

$$c' \stackrel{\text{def}}{=} \frac{c}{1-\varepsilon}.$$

To get the inequality related to the probability $p = 2^{-n}$, we need to apply the function 2^{-x} to both sides of the above inequality, resulting in

$$p < 2^{-a \cdot K(x)} \cdot p_0, \tag{1}$$

where we denoted $a \stackrel{\text{def}}{=} 1/(1-\varepsilon)$ and $p_0 \stackrel{\text{def}}{=} 2^{-c}$.

Towards the resulting consistent formalization of the principle that events with small probability cannot happen. The above conclusion was made only for the probability measure that corresponds to flipping a coin, when all elements in a sequence are independent and each is equal to 0 or 1 (or T or H) with equal probability 0.5. However, we can repeat the same arguments for all other probability distributions – especially taking into account that all reasonable continuous probability distributions are in 1-1-correspondence with each other.

Thus, we arrive at the following definition.

4.4 Algorithmic Information Theory approach to hypothesis testing

Definition 4. *Let X be a definable set, let p be a definable probability measure on X, and let ε and $p_0 \ll 1$ be positive numbers. We say that an element x is* (c, ε)-random *if it does not belong to any definable set E for which*

$$p(E) \le 2^{-a \cdot K(E)} \cdot p_0, \tag{1}$$

where we denoted $p_0 = 2^{-c/(1-\varepsilon)}$ and $a = 1/(1-\varepsilon)$.

Discussion.

- So, if we observe an event whose probability, according to a given hypothesis, is too small, this means that this hypothesis must be rejected.
- Our inequality (1) is similar to the p-value-type inequality $p(E) \le p_0$ that caused many troubles. However, there is now a big difference: the threshold probability for rejecting a hypothesis is no longer a constant, it changes depending on the complexity $K(E)$ of formulating the hypothesis. Let us show that this indeed solves all the crisis situations.

A mathematical comment: this solution is related to the Bayesian approach. In the Bayesian approach, we start with a prior distribution – that reflects all the knowledge that we have so far. As we gain new knowledge, we use the Bayes formula to update this distribution.

An important question is: what prior distribution m_0 should we select if we do not have any prior knowledge at all? In general, it should not matter that much: it is

known that under some reasonable conditions, then, no matter what prior distribution we start with, eventually, with more and more observations, we get closer and closer to the actual distribution. The "reasonable conditions" means, in effect, that if we have an event for which the actual probability is positive, then the prior probability of this event should also be positive. Otherwise, if for some event, the prior probability is 0, then, no matter how many observations we get, this probability will remain to be 0 – and it will never get positive.

In situations in which we have no information about the actual distribution – so that any definable distribution is possible – we should select a prior distribution for which for any set E, if $p(E) > 0$ for some definable probability distribution, then we should have $m_0(E) > 0$. It turns out – see, e.g., [10] – that this condition implies that $m_0(E)$ is equal to $2^{-K(E)}$ for some version of a formal language.

So, in terms of this prior distribution, the Algorithmic Information Theory approach to hypothesis testing can be formulated as follows: an element is random if it does not belong to any set E for which $p(E) \le m_0(E) \cdot p_0$. In other words, we eliminate as impossible all the events E for which the actual probability is much smaller than the prior probability, i.e., for which $p(E)/m_0(E) \le p_0$ for some small p_0.

4.5 The new approach solves the foundational crisis: the new approach is consistent

Proposition 2. *Let X be a definable set, and let p be a definable probability measure on X. Then, when $p_0 < 2^{\varepsilon/(1-\varepsilon)} - 1$, then the set R of all random elements has a positive measure*

$$m(R) \ge 1 - \frac{p_0}{2^{\varepsilon/(1-\varepsilon)} - 1}.$$

Discussion.

- In other words, the new definition of randomness is consistent.
- For small $\varepsilon \ll 1$, we have $1 - \varepsilon \approx 1$, so $\varepsilon/(1-\varepsilon) \approx \varepsilon$, and $2^{\varepsilon} - 1 = \exp(\ln 2 \cdot \varepsilon) - 1 \approx \ln(2) \cdot \varepsilon$. Thus, for small ε, the inequality from Proposition 2 becomes, approximately, $p_0 < \ln 2 \cdot \varepsilon$, and the probability of an element to be random is approximately equal to
$$1 - \frac{p_0}{\ln(2) \cdot \varepsilon}.$$

Proof. Kolmorogov complexity $K(E)$ is the smallest binary length of the statement that uniquely describes this set. For each $n \ge 1$, there are 2^n binary sequences of length n and thus, at most 2^n sets E for which $K(E) = n$. So, the overall measure of all the sets E with $K(E) = n$ that are rejected because their probability is small is smaller than or equal to the sum of their probabilities. For each of these rejected

sets, the probability is smaller than $2^{-a\cdot n}\cdot p_0$, Thus, the overall measure of all $\le 2^n$ such rejected sets cannot exceed $2^n\cdot 2^{-a\cdot n}\cdot p_0 = 2^{-(a-1)\cdot n}\cdot p_0$.

The overall measure r of all rejected sets, with all possible values $n = K(E)$, cannot exceed the sum of the measures rejected for each n:

$$r \le \sum_{n=1}^{\infty} 2^{-(a-1)\cdot n}\cdot p_0 = p_0\cdot \sum_{n=1}^{\infty}\left(2^{-(a-1)}\right)^n.$$

The sum on the right-hand side is a geometric progression, so its sum is determined by a known formula

$$s \stackrel{\text{def}}{=} \sum_{n=1}^{\infty}\left(2^{-(a-1)}\right)^n = \frac{2^{-(a-1)}}{1-2^{-(a-1)}}.$$

If we multiply both numerator and generator by 2^{a-1}, we get a simplified expression

$$s = \frac{1}{2^{a-1}-1}.$$

Here,

$$a-1 = \frac{1}{1-\varepsilon}-1 = \frac{\varepsilon}{1-\varepsilon},$$

so

$$s = \frac{1}{2^{\varepsilon/(1-\varepsilon)}-1},$$

and thus,

$$r \le s\cdot p_0 = \frac{p_0}{2^{\varepsilon/(1-\varepsilon)}-1}.$$

Thus, the measure of the set of all random elements is at least as large as 1 minus r.

The proposition is proven.

4.6 The new approach helps to solve the reproducibility crisis

As we have mentioned, one of the main reasons for the reproducibility crisis is the general idea behind the p-value approach: that events with probability $p \le p_0$ are not possible. For $p_0 = 0.05$, this means that, on average, one of 20 randomly selected hypotheses will satisfy this inequality – and thus, will be mistakenly claimed to be true.

What will happen if we try to apply the same technique in the new setting? For every length n, there are only finitely many hypotheses of Kolmogorov complexity $K(E) = n$. Thus, if we try to test many hypotheses, we will inevitably exhaust this length and thus, have to consider hypotheses for which the Kolmogorov complexity is higher – and thus, the threshold $2^{-K(E)}\cdot p_0$ is even smaller. Since the probability is smaller, to get the same p-hacking result as before, we need even more hypotheses

and tests – which will bring us to cases with an even larger $K(E)$ and thus, even smaller threshold $2^{-K(E)} \cdot p_0$, etc.

4.7 The new approach helps to solve the crisis of adequacy

Crisis: a brief reminder. As we have mentioned, the crisis of adequacy is that:

- while physicists believe that events with very small probability cannot happen – e.g., atoms in a gas cannot, by themselves, form the text "Hi from atoms",
- a usual statistical approach claims that all events are possible, even events with very small probability.

The new approach helps bridge the gap between scientific researchers and statistical techniques. In the new approach, we explicitly reject some starting sequences as not possible for a random event – in perfect agreement with the scientists' viewpoint.

5 Practical considerations and remaining open problems

5.1 What we do in this section

The above solution is good from the theoretical viewpoint, in the sense that it helps resolve the three crises that we started with. However, in order to use this approach in practice, we need to be able to answer several questions:

- which of several possible functions $K(x)$ should we use?
- how can we use this approach – taking into account that Kolmogorov complexity is, in general, not computable?, and
- what should we do if we do not have enough data to reject a hypothesis?

Let us summarize what is known about these questions. For some of these questions, we have partial answers; other questions remains open.

5.2 Which of the possible functions $K(x)$ should we use?

Question. As we have mentioned, the value of the Kolmogorov complexity depends on what formal language we select to represent different properties. Which of these values should we use?

Who should decide on this? We are not talking about abstract mathematics here, we are talking about which events are actually possible and which are not – i.e., we

are talking about the real world. So, the answer to this question should come from the science that studies the real world – i.e., from physics.

Usually, physicists simply provide the probabilities of different events, not realizing that for the same probability distribution, we may have different conclusions about what is possible and what is not. This needs to be studied experimentally and theoretically.

Maybe we should go in an even more general direction? Since we are talking about the physical world, it is not even clear that the physical world should follow one of our formal systems. Maybe we should look for a more general idea, not related to any specific formal system.

Such an attempt is described, e.g., in [3, 5, 6, 7, 8, 9], where the idea that events with small probability cannot happen is formalized as follows: for every definable nested sequence of sets $S_1 \supset S_2 \supset \ldots$ for which $p(S_n) \to 0$, there exists an i for which all the elements in set S_i are not random.

On the qualitative level, this general approach has all the advantages of an approach based on Kolmogorov complexity. However, from the practical viewpoint, it is even less clear how to apply this general approach.

5.3 Once we fix $K(x)$, how can we use this approach – taking into account that Kolmogorov complexity is, in general, not computable

Problem. Suppose that we have answered the first question, and we know which version of $K(x)$ we should use. Now, we face another question, related to the known fact that $K(x)$ is, in general, not computable (see, e.g., [10]): How can we check the inequality $p(E) < 2^{-K(E)} \cdot p_0$?

A natural idea. A natural idea is to use some *approximation* to the Kolmogorov complexity $K(E)$. For example, since $K(E)$ is the shortest possible description of the set E, for any other description, its length ℓ is greater than or equal to $K(E)$. In this case, $2^{-\ell} \le 2^{-K(E)}$.

So, if the above inequality holds with the approximate value ℓ instead of $K(E)$, i.e., if we have $p(E) \le 2^{-\ell} \cdot p_0$, then automatically $p(E) < 2^{-K(E)} \cdot p_0$ and thus, event E is impossible – and the corresponding hypothesis can be rejected.

5.4 What if we do not have enough data to reject a hypothesis with 100% certainty

Question. In situations like physics, where we can perform many experiments, we may be able to reach a small number p_0 – and thus, absolutely reject a hypothesis.

However, in economics, in psychology, in social sciences, the number of possible observations and experiments is limited. In these sciences, we may not be able to reach the absolute rejection level. What should we do in this case?

A possible solution. In this case, what we can do is to use the new criterion, but with larger values p_0 – e.g., $p_0 = 0.05$ or $p_0 = 0.001$.

When we "reject" a hypothesis based on such value p_0, it does not necessarily mean that this hypothesis is false, it simply means that there is a certain degree of confidence that it is false.

Comment. This is similar to the usual p-value approach. The main difference is that now, our approach is consistent.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI), by the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany, and by the European Union under the project ROBOPROX (No. CZ.02.01.01/00/22 008/0004590).

This work also was supported by the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand.

References

1. Y. Benjamini, R. D. De Veaux, B. Efron, S. Evans, M. Glickman, B. T. Graubard, X. He, X.-L. Meng, N. M. Reid, S. M. Stigler, S. B. Vardeman, C. K. Wikle, T. Wright, L. J. Young, and K. Kafadar, "ASA President's Task Force Statement on Statistical Significance and Replicability", *Chance*, 2021, Vol. 34, No. 4, pp. 10–11. doi:10.1080/09332480.2021.2003631
2. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
3. A. M. Finkelstein and V. Kreinovich, "Impossibility of hardly possible events: physical consequences," *Abstracts of the 8th International Congress on Logic, Methodology, and Philosophy of Science*, Moscow, 1987, Vol. 5, Part 22, pp. 23–25.
4. J. P. Ioannidis, "Why most published research findings are false", *PLOS Medicine*, 2005, Vol. 2, No. 8, Paper e124, doi:10.1371/journal.pmed.0020124
5. V. Kreinovich, "Toward formalizing non-monotonic reasoning in physics: the use of Kolmogorov complexity and Algorithmic Information Theory to formalize the notions 'typically' and 'normally' ", In: L. Sheremetov and M. Alvarado (eds.), *Proceedings of the Workshops on Intelligent Computing WIC'04 associated with the Mexican International Conference on Artificial Intelligence MICAI'04*, Mexico City, Mexico, April 26–27, 2004, pp. 187–194.

6. V. Kreinovich, "Toward formalizing non-monotonic reasoning in physics: the use of Kolmogorov complexity", *Revista Iberoamericana de Inteligencia Artificial*, 2009, Vol. 41, pp. 4–20.
7. V. Kreinovich, "Towards formalizing non-monotonic reasoning in physics: logical approach based on physical induction and its relation to Kolmogorov complexity", in: E. Erdem, J. Lee, Y. Lierler, and D. Pearce (eds.), *Correct Reasoning: Essays on Logic-Based AI in Honor of Vladimir Lifschitz*, Springer Lecture Notes on Computer Science, 2012, Vol. 7265, pp. 390–404.
8. V. Kreinovich and A. M. Finkelstein, "Towards applying computational complexity to foundations of physics", *Notes of Mathematical Seminars of St. Petersburg Department of Steklov Institute of Mathematics*, 2004, Vol. 316, pp. 63–110; reprinted in *Journal of Mathematical Sciences*, 2006, Vol. 134, No. 5, pp. 2358–2382.
9. V. Kreinovich, L. Longpré, and H. T. Nguyen, "Towards formalization of feasibility, randomness, and commonsense implication: Kolmogorov complexity, and the necessity of considering (fuzzy) degrees", In: N. H. Phuong and A. Ohsato (eds.), *Proceedings of the Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications VJFUZZY'98*, HaLong Bay, Vietnam, 30th September-2nd October, 1998, pp. 294–302.
10. M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, Cham, Switzerland, 2019.
11. K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2021.
12. R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-values: context, process, and purpose", *The American Statistician*, 2016, Vol. 70, No. 2, pp. 129–133, doi:10.1080/00031305.2016.1154108.