Use of the fundamental problem of statistics to define the validity limit of Occam's razor principle

Andrea Berdondini

ABSTRACT: In statistics, to evaluate the significance of a result, one of the most used methods is the statistical hypothesis test. Using this theory, the fundamental problem of statistics can be expressed as follows: "*A statistical data does not represent useful information, but becomes useful information only when it is shown that it was not obtained randomly*". Consequently, according to this point of view, among the hypotheses that perform the same prediction, we must choose the result that has a lower probability of being produced randomly. Therefore, the fundamental aspect of this approach is to calculate correctly this probability value. This problem is addressed by redefining what is meant by hypothesis. The traditional approach considers the hypothesis as the set of rules that actively participate in the forecast. Instead, we consider as hypotheses the sum of all the hypotheses made, also considering the hypotheses preceding the one used. Therefore, each time a prediction is made, our hypothesis increases in complexity and consequently increases its ability to adapt to a random data set. In this way, the complexity of a hypothesis can be precisely determined only if all previous attempts are known. Consequently, Occam's razor principle no longer has a general value, but its application depends on the information we have on the tested hypotheses.

Introduction

The logical principle of Occam's razor [1], [2], suggests choosing the simplest hypothesis among those available. In this article, we will analyze this principle using the theory of statistical hypothesis test [3], [4]. By exploiting this theory, we will reformulate the fundamental problem of statistics in such a way as to bring attention to the link between the statistical data and the probability that it was produced randomly. Consequently, according to this point of view, among the hypotheses that perform the same prediction, we must choose the result that has a lower probability of being produced randomly. Therefore, it becomes essential to calculate this probability value correctly.

This problem is addressed by redefining what is meant by hypothesis. The traditional approach considers the hypothesis as the set of rules that actively participate in the forecast. Instead, we consider as hypotheses the sum of all the hypotheses made, also considering the hypotheses preceding the one used. Therefore, each time a prediction is made, our hypothesis increases in complexity and consequently increases its ability to adapt to a random data set. In this way, the complexity of a hypothesis can be precisely determined only if all previous attempts are known. Consequently, Occam's razor principle no longer has a general value, but its application depends on the information we have on the tested hypotheses.

Finally, we use this new definition of hypothesis to understand the reason for the high percentage of non-reproducible results, in which the hypothesis test was used.

The fundamental problem of statistics

In statistics, to evaluate the significance of a result, one of the most used methods is the statistical hypothesis test. Using this theory, the fundamental problem of statistics can be expressed as follows: "*A statistical data does not represent useful information, but becomes useful information*

only when it is shown that it was not obtained randomly".

 This definition is particularly significant, because it highlights the two fundamental aspects of statistics, which are its uncertainty and the reason for its uncertainty. Indeed, the purpose of statistics is the study of phenomena in conditions of uncertainty or non-determinism by exploiting the sampling of events related to the phenomenon to be studied. Knowing that the observed events can be randomly reproduced with a probability that will never be zero, we understand the reason for the indeterminism that characterizes the statistics. This probability value is called universal probability [5].

 Through this definition of the fundamental problem of statistics, it is also possible to formulate the following paradox [6], which highlights how the evaluation of statistical results is dependent on each action performed on the analyzed data.

The validity limit of Occam's razor principle

 In this paragraph, we will see how the information regarding the development of a hypothesis is fundamental to define the validity limit of Occam's razor principle.

 Let us start by giving some definitions useful to formalize our theory.

 Given an experiment that measures N values of a discrete variable X with cardinality C, we call D the set of dimension C^N, which includes all possible sequences X^N of length N that can be observed.

 Now, we redefine the concept of hypothesis in order to define a chronological succession among the tested hypothesis.

 We call *H(t)* the hypothesis developed at time *t*.

 We call *PH(t)* the set of sequences $X^N \in D$ that the hypothesis *H(t)* is able to predict.

 We call *NPH(t)* the cardinality of the set *PH(t)*.

 We call *TH(t)* the set that includes all the hypotheses up to time *t*.

$$TH(t) = \{H(i_1), H(i_2), \dots \dots, H(i_t)\}$$

 We call *TPH(t)* the union of all the sets *PH(t)* relating to all the hypotheses *H(t)* $\in$ *TH(t)*.

$$TPH(t) = \bigcup_{i=0}^{t} PH(i)$$

 We call *NTPH(t)* the cardinality of the set *TPH(t)*. Consequently, *NTPH(T)* defines the number of sequences, belonging to *D,* that the hypothesis *TH(t)* is able to predict. It may happen that different hypotheses forecast the same sequence of values of *X*, having made the union of the sets *PH(t)* these sequences are calculated only once.

 If we have only made a hypothesis *H(t)=TH(t)* and *NPH(t)=NTPH(t)*.

If, on the other hand, more than one hypothesis has been tested $H(t) \neq TH(t)$ and $NPH(t) \leq NTPH(t)$.

We define the ability of the hypothesis $TH(t)$ to predict a sequence of N casual observations of the i.i.d. random variable X with discrete uniform distribution and cardinality C, the ratio:

$$\frac{NTPH(t)}{C^N} \tag{1}$$

This ratio also defines the probability that the hypothesis $TH(t)$ can predict the results of an experiment, in which the cardinality of D is equal to C^N, in a completely random way.

Knowing that a hypothesis $TH(t)$ can predict the results of an experiment only in the following two conditions:

1) $TH(t)$ is true.

2) $TH(t)$ is false and the prediction occurs randomly; the probability of this event is given by equation (1).

Under these conditions, the probability that the hypothesis $TH(t)$ is true turns out to be:

$$1 - \frac{NTPH(t)}{C^N} \tag{2}$$

Consequently, this equation defines the parameter that must be used in the evaluation of $H(t)$. So, if we want to compare two hypotheses $H1(t)$ and $H2(t)$, we have 4 possible results:

1) $NPH1(t)>NPH2(t)$ e $NTPH1(t)>NTPH2(t)$

2) $NPH1(t)>NPH2(t)$ e $NTPH1(t)<NTPH2(t)$

3) $NPH1(t)<NPH2(t)$ e $NTPH1(t)<NTPH2(t)$

4) $NPH1(t)<NPH2(t)$ e $NTPH1(t)>NTPH2(t)$

$NPH(t)$ and $NTPH(t)$ define the number of sequences that hypothesis $H(t)$ and the hypothesis $TH(t)$ are able to predict. Consequently, they can be used as a measure of their complexity, in fact, the more complex a hypothesis is, the greater the number of results it can predict.

Analyzing the four possible results, we note that even if a hypothesis $H1(t)$ is less complex than a hypothesis $H2(t)$ ($NPH1(t)<NPH2(t)$), it is possible to have a hypothesis $TH1(t)$ more complex than a hypothesis $TH2(t)$ ($NTPH1(t)>NTPH2(t)$). Consequently, using equation (2) as an evaluation method, hypothesis $H1(t)$ should be discarded in favor of hypothesis $H2(t)$. This situation can happen, for example, if $H1(t)$ is the last hypothesis of a long series of other hypotheses tested previously.

In the event that there is no information on the hypotheses to be evaluated, it must be assumed that the hypotheses have been developed under the same conditions. Therefore, in this case, not being able to calculate $TH(t)$, it is recommended to choose the simpler hypothesis $H(t)$.

Finally, from equation (2), we can deduce the following result: given a hypothesis $H(t)$ the probability that is true can be calculated only if all the previously tested hypotheses are known.

Consequently, the complexity of a hypothesis not only depends on the mathematical formula that makes the prediction, but also depends on all the attempts made previously. Therefore, Occam's razor principle does not have an absolute value but its application depends on the information about the hypotheses.

How to perform correctly the statistical hypothesis test

It is interesting to note how the definition of hypothesis, which was given in the previous paragraph, can be seen as something extremely obvious or as something extremely innovative. Indeed, it may seem absolutely banal to consider all the hypotheses that have been tested, for the obvious reason that by running a large number of random hypotheses sooner or later there will be some hypothesis that will fit the data quite well. On the other hand, also considering the previous hypotheses represents a revolution in the evaluation of a hypothesis. In fact, from this point of view, the mere knowledge of the hypothesis that makes the prediction does not allow us to define its real complexity.

Therefore, if in the statistical hypothesis test the p-value [7], [8], used as a threshold to reject the null hypothesis, is calculated considering only the hypothesis that actively participates in the prediction, it means, that we are underestimating the complexity of the hypothesis. Consequently, the p-value, thus calculated, is wrong and therefore determines a false evaluation of the hypothesis. It is therefore believed that this systematic error, in the execution of the hypothesis test, is responsible for the high number of non-reproducible results [9], [10].

Taking advantage of these considerations it is understood that evaluating a statistical result can be very difficult because some information can be hidden. For example, we are obliged to report the mathematical formula that makes the prediction but, instead, we may not report all previous failed attempts. Unfortunately, this information is essential for evaluating the hypothesis, because they are an integral part of the hypothesis. Indeed, if we test 10 hypotheses, we simply interpolate the data with those ten hypotheses and choose the hypothesis that passes the chosen evaluation test.

This problem also depends on the increasing use of statistical software capable of quickly executing a huge number of mathematical models. Consequently, there is the risk of "playing" with this software by performing a multitude of analyzes and this sooner or later leads to a random correlation.

For these reasons, the evaluation of statistical results represents one of the most important challenges for scientific research. Unfortunately, it is a difficult problem to solve because, as mentioned, some information can always be hidden when writing an article. The simplest solution adopted is to use more selective evaluation parameters, which in practice means making it unlikely to pass the evaluation test by developing random hypotheses. However, this solution has different problems in fact, in this way, we can discard correct hypotheses and cannot be applied to all fields of research. For example, in finance where the possible inefficiencies of the markets [11], which can be observed, are minimal, adopting very restrictive valuation methods means having to discard almost any hypothesis.

Conclusion

In this article, we have discussed the logical principle of Occam's razor using the hypothesis test theory. This allowed us to reformulate the fundamental problem of statistics, in such a way as to make us understand the importance of correctly calculating the probability of obtaining the same results randomly. Solving this problem involved redefining the concept of hypothesis. According to this point of view, by hypothesis we mean the sum of all tested hypotheses. Consequently, the complexity of a hypothesis not only depends on the mathematical formula that makes the prediction but also depends on the previous hypotheses tested.

Therefore, according to this approach, the logical principle of Occam's razor no longer has a general value if one considers as a hypothesis only the set of rules that actively participate in the prediction. If, on the other hand, the hypothesis is considered as the sum of all the tested hypotheses, in this case, Occam's razor principle returns to have a general value.

Finally, it is noted that not considering all the tested hypotheses causes a systematic error in the application of the statistical hypothesis test. Therefore, it is hypothesized that this error, which leads to underestimate the complexity of a hypothesis, is the cause of the high percentage of non-reproducible scientific results.

References

[1] Roger Ariew, "Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony", 1976.

[2] Sober, Elliott (2004). "What is the Problem of Simplicity?". In Zellner, Arnold; Keuzenkamp, Hugo A.; McAleer, Michael (eds.). Simplicity, Inference and Modeling: Keeping it Sophisticatedly Simple. Cambridge, U.K.: Cambridge University Press. pp. 13–31. ISBN 978-0-521-80361-8. Retrieved 4 August 2012ISBN 0-511-00748-5 (eBook [Adobe Reader]) paper as pdf.

[3] Fisher, R (1955). "Statistical Methods and Scientific Induction" (PDF). Journal of the Royal Statistical Society, Series B. 17 (1): 69–78.

[4] Borror, Connie M. (2009). "Statistical decision making". The Certified Quality Engineer Handbook (3rd ed.). Milwaukee, WI: ASQ Quality Press. pp. 418–472. ISBN 978-0-873-89745-7.

[5] Cristian S. Calude (2002). "Information and Randomness: An Algorithmic Perspective", second edition. Springer. ISBN 3-540-43466-6.

[6] Berdondini Andrea, "The Information Paradox", (July 8, 2019). Available at SSRN: https://ssrn.com/abstract=3416559.

[7] Wasserstein, Ronald L.; Lazar, Nicole A. (7 March 2016). "The ASA's Statement on p-Values: Context, Process, and Purpose". The American Statistician. 70 (2): 129–133. doi:10.1080/00031305.2016.1154108.

[8] Hung, H.M.J.; O'Neill, R.T.; Bauer, P.; Kohne, K. (1997). "The behavior of the p-value when the alternative hypothesis is true". Biometrics (Submitted manuscript). 53 (1): 11–22.

[9] Munafò, M., Nosek, B., Bishop, D. et al. "A manifesto for reproducible science". Nat Hum Behav 1, 0021 (2017). https://doi.org/10.1038/s41562-016-0021.

[10] Ioannidis, J. P. A. "Why most published research findings are false". PLoS Med. 2, e124 (2005).

[11] Black, F. (1971) "Random Walk and Portfolio Management," Financial Analyst Journal, 27, 16-22.

E-mail address: andrea.berdondini@libero.it