The information paradox

Andrea Berdondini

ABSTRACT: The following paradox is based on the consideration that the value of a statistical datum does not represent a useful information, but becomes a useful information only when it is possible to proof that it was not obtained in a random way. In practice, the probability of obtaining the same result randomly must be very low in order to consider the result useful. It follows that the value of a statistical datum is something absolute but its evaluation in order to understand whether it is useful or not is something of relative depending on the actions that have been performed. So two people who analyze the same event, under the same conditions, performing two different procedures obviously find the same value, regarding a statistical parameter, but the evaluation on the importance of the data obtained will be different because it depends on the procedure used. This condition can create a situation like the one described in this paradox, where in one case it is practically certain that the statistical datum is useful, instead in the other case the statistical datum turns out to be completely devoid of value. This paradox wants to bring attention to the importance of the procedure used to extract statistical information; in fact the way in which we act affects the probability of obtaining the same result in a random way and consequently on the evaluation of the statistical parameter.

The information paradox

We have two identical universes, in both universes the same person is present, that we will call John, he must perform the exact same task which is to analyze a database in order to extract useful correlations. As we have said the universes are equal, so the databases are identical and the person who has to do the work is the same. The database that needs to be analyzed consists of a million parameters related to an event to be studied.

In the universe "1”, John acts as follows: he takes the whole database and calculates the correlation of the parameters with the event to be studied. From this analysis he finds 50 parameters with a high correlation with the event, the correlation found has a chance to happen randomly of 0.005%. Of these 50 parameters, John identifies 10 that according to his experience can be useful in order to study the event. However it is important to point out that the assumptions made by John, on the 10 parameters, are only hypotheses based on his experience, they are not scientific demonstrations that explain precisely the correlation of the 10 parameters with the event.

In the universe "2", John acts in the following way: before analyzing the entire database he uses his knowledge of the event in order to select 10 parameters, that he believes are most correlated with the event, from the million parameters available. However, also in this case, it is important to point out that the assumptions made by John, on the 10 parameters, are only hypotheses based on his experience, they are not scientific demonstrations that explain precisely the correlation of the 10 parameters with the event. Analyzing only these 10 parameters, he finds 5 of them with a high correlation with the event, the correlation found has a chance to happen randomly of 0.005% (as in the previous case).

In practice, the fundamental difference in the analysis method that John does in the two universes is that: in the first universe John uses his own experience after performing statistical analysis on the whole database, instead in the second universe, John uses his experience before to perform the statistical analysis in order to reduce the size of the database.

Now let us see how this different approach affects the evaluation of the data obtained. To do this, we must calculate the probability of obtaining the same results randomly in the two cases.

In the first case, universe "1", in order to calculate the probability of obtaining the same results in a random way we must use the binomial distribution formula with the following parameters:

probability of victory (p) = probability of getting the same correlation randomly

number of successes (k) = number of parameters that present the correlation considered

number of tests (L) = total number of parameters present in the database

By entering these data within the binomial distribution formula:

$$F(k, L, p) = \binom{L}{k} p^k (1-p)^{L-k}$$

p = 0.005%

k = 50

L = 1 Million

We get a probability of 5.6% as a result.

Now let's consider the second case, the universe "2", even in this situation, in order to calculate the probability of obtaining the same results in a random way we must use the binomial distribution formula with the following parameters:

p = 0.005%

k = 5

L = 10

The probability obtained in this case is $7.9 \ \cdot 10^{-18}$ %.

Analyzing these results it is easy to understand that a percentage of 5.6% makes the correlations found not significant. In order to understand how high this percentage is, we can also calculate the probability of obtaining, in a random way, more than 50 of parameters with the correlation considered, this probability is 46%.

Now we analyze the percentage of the second case ($7.9 \ \cdot 10^{-18}$ %) this percentage is extremely low, consequently we are practically certain that the correlation found is not random and therefore this result represents a useful information for studying the event.

At this point, John must to decide whether to implement the correlations found or not. Obviously, exploiting the correlations found implies costs, therefore a wrong evaluation involves a high risk. In the universe “1” John is in a difficult situation, in fact the work done is not only useless but also dangerous because it can lead him to sustain wrong investments. Instead, in the second universe John knows that the probability that the correlation is random is almost zero, so he can invest with an acceptable risk.

In conclusion, a simple procedural error has led to enormous consequences. In the first case the experience of john is useless, instead in the second case it was a key resource in order to extract useful information from a big database.

In fact, in the case of the universe “1”, John can no longer use his own knowledge and the only thing he can do is transform his hypotheses into real scientific demonstrations, but in many situations, as in the financial field, doing it can be very difficult. Consequently, when hypotheses are made after having carried out an analysis, these hypotheses risk being conditioned by the results and therefore lose value. Instead, the hypotheses made before the analysis are not conditioned and the analysis of the data is used in order to verify them in a statistical way, as happened in the universe “2”.

One of the fields where it is fundamental to calculate the probability of obtaining the same data in a random way, as a method of evaluating the correlations detected, is the financial one [1], [2].

Conclusion

In this article we have used a paradox to explain how a statistical datum does not represent a useful information, it becomes a useful information, to study an event, only when it is possible to prove that the probability that it was obtained in a random way is very low. This consideration makes the application of statistics, as a method of evaluating a hypothesis, a "relativistic" science. In fact, as described in the paradox, the calculation of the probability of obtaining the same result in a random way is something of relative that depend from the method used and from the actions performed.

These considerations have a great impact from an experimental point of view, because they teach us the importance of correct planning, in which we must always implement all the knowledge about the event we want to study. It is also essential keep track of all the operations performed on the data, because this information is necessary in order to calculate correctly the probability of obtaining the same results in a random way.

This way of interpreting statistical data is also very useful for understanding the phenomenon of overfitting, a very important issue in data analysis [3], [4]. The overfitting seen from this point of view is simply the direct consequence of considering the statistical parameters, and therefore the results obtained, as a useful information without checking that them was not obtained in a random way. Therefore, in order to estimate the presence of overfitting we have to use the algorithm on a database equivalent to the real one but with randomly generated values, repeating this operation many times we can estimate the probability of obtaining equal or better results in a random way. If this probability is high, we are most likely in an overfitting situation. For example, the probability that a fourth-degree polynomial has a correlation of 1 with 5 random points on a plane is 100%, so this correlation is useless and we are in an overfitting situation.

This approach can also be applied to the St Petersburg paradox [5], in fact also in this case the expectation gain is a statistical datum that must be evaluated before being used at the decisional level. In fact, the difficulty in solving the paradox stems from the fact of considering a statistical datum always as a useful information. Analyzing the expectation gain it is possible to proof that we can obtain better result, randomly, with a probability that tends asymptotically to 50%. Consequently, the expectation gain that tends to infinity turns out to be a statistic data without value that cannot be used for decision-making purposes.

This way of thinking gives an explanation to the logical principle of Occam's razor, in which it is advisable to choose the simplest solution among the available solutions. In fact, for example, if we want to analyze some points on a plane with a polynomial, increasing the degree increases the probability that a given correlation can occur randomly. For example, given 24 points on a plane, a second degree polynomial has a 50% probability of randomly having a correlation greater than 0.27, instead a fourth degree polynomial has a probability of 84% of having a correlation greater than 0.27 randomly. Therefore, the value of the correlation is an absolute datum but its validity to study a set of data is something relative that depends on the method used. Consequently the simpler methods, being less parameterized, have a lower probability of a randomly correlation, so they are preferred over the complex methods.

References

[1] Andrea Berdondini, "Application of the Von Mises' Axiom of Randomness on the Forecasts Concerning the Dynamics of a Non-Stationary System Described by a Numerical Sequence" (January 21, 2019). Available at SSRN: https://ssrn.com/abstract=3319864 or http://dx.doi.org/10.2139/ssrn.3319864.

[2] Andrea Berdondini, "Description of a Methodology from Econophysics as a Verification Technique for a Financial Strategy", (May 1, 2017). Available at SSRN: https://ssrn.com/abstract=3184781.

[3] Igor V. Tetko, David J. Livingstone, and Alexander I. Luik, "Neural network studies. 1. Comparison of overfitting and overtraining", Journal of Chemical Information and Computer Sciences 1995 35 (5), 826-833 DOI: 10.1021/ci00027a006.

[4] Quinlan, J.R. (1986). "The effect of noise on concept learning". In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.),Machine learning: An artificial intelligence approach(Vol. 2). San Mateo, CA: Morgan Kaufmann.

[5] Andrea Berdondini, "Resolution of the St. Petersburg Paradox Using Von Mises' Axiom of Randomness" (June 3, 2019). Available at SSRN: https://ssrn.com/abstract=3398208.

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