# The Fundamental Problem of Causal Inference

Andrea Berdondini

ABSTRACT: The fundamental problem of causal inference defines the impossibility of associating a causal link to a correlation, in other words: correlation does not prove causality. This problem can be understood from two points of view: experimental and statistical. The experimental approach tells us that this problem arises from the impossibility of simultaneously observing an event both in the presence and absence of a hypothesis. The statistical approach, on the other hand, suggests that this problem stems from the error of treating tested hypotheses as independent of each other. Modern statistics tends to place greater emphasis on the statistical approach because, compared to the experimental point of view, it also shows us a way to solve the problem. Indeed, when testing many hypotheses, a composite hypothesis is constructed that tends to cover the entire solution space. Consequently, the composite hypothesis can be fitted to any data set by generating a random correlation. Furthermore, the probability that the correlation is random is equal to the probability of obtaining the same result by generating an equivalent number of random hypotheses.

## Introduction

The fundamental problem of causal inference defines the impossibility of associating causality with a correlation; in other words, correlation does not prove causality. This problem can be understood from two perspectives: experimental and statistical. The experimental approach suggests that this problem arises from the impossibility of observing an event both in the presence and absence of a hypothesis simultaneously. The statistical approach, on the other hand, suggests that this problem stems from the error of treating tested hypotheses as independent of each other.

Modern statistics tends to place greater emphasis on the statistical approach, as it, unlike the experimental approach, also provides a path to solving the problem. Indeed, when testing many hypotheses, a composite hypothesis is constructed that tends to cover the entire solution space. Consequently, the composite hypothesis can fit any data series, thereby generating a correlation that does not imply causality.

Furthermore, the probability that the correlation is random is equal to the probability of obtaining the same result by generating an equivalent number of random hypotheses. Regarding this topic, we will see that the key point, in calculating this probability value, is to consider hypotheses as dependent on all other previously tested hypotheses.

Considering the hypothesis as non-independent has fundamental implications in statistical analysis. **In fact, every random action we take is not only useless but will increase the probability of a random correlation.** For this reason, in the following article [1], we highlight the importance of acting consciously in statistics.

Moreover, calculating the probability that the correlation is random is only possible if all prior attempts are known. In practice, calculating this probability is very difficult because not only do we need to consider all our attempts, but we also need to consider the attempts made by everyone else performing the same task. In fact, a group of people belonging to a research network all having the same reputation and all working on the same problem can be considered with a single person who performs all the attempts made. From a practical point of view, we are almost always in the situation

where this parameter is underestimated because it is very difficult to know all the hypotheses tested. Consequently, the calculation of the probability that a correlation is casual becomes something relative that depends on the information we have.

## The Fundamental Problem of Causal Inference

The fundamental problem of causal inference [2] defines the impossibility of associating causality with a correlation, in other words: correlation does not prove causality. From a statistical point of view, this indeterminacy arises from the error of considering the tested hypotheses as independent of each other. When a series of hypotheses is generated, a composite hypothesis is formed that tends to fit any data series, leading to purely random correlations.

For example, you can find amusing correlations between very different events on the internet; these correlations are obviously random. These examples are often used to demonstrate the fundamental problem of causal inference**. In presenting this data, the following information is always omitted: how many hypotheses did I consider before finding a related hypothesis.**

This is essential information because if I have a database comprising a very high number of events, for any data series, there will always be a hypothesis that correlates well with my data. Thus, if I generate a large number of random hypotheses, I will almost certainly find a hypothesis that correlates with the data I am studying. Therefore, having a probability of about 100% of being able to obtain the same result randomly, I have a probability of about 100% that the correlation does not also imply causation.

On the other hand, if we generate a single hypothesis that correlates well with the data, in this situation, almost certainly, the correlation also implies causation. This is because the probability of obtaining a good correlation by generating a single random hypothesis is almost zero.

This result is also intuitive, because it is possible to achieve a good correlation with a single attempt only if one has knowledge of the process that generated the data to be analyzed. And it is precisely this knowledge that also determines a causal constraint.

Figure 1 summarizes the basic concepts showing how the correct way to proceed is to consider the hypotheses as non-independent.

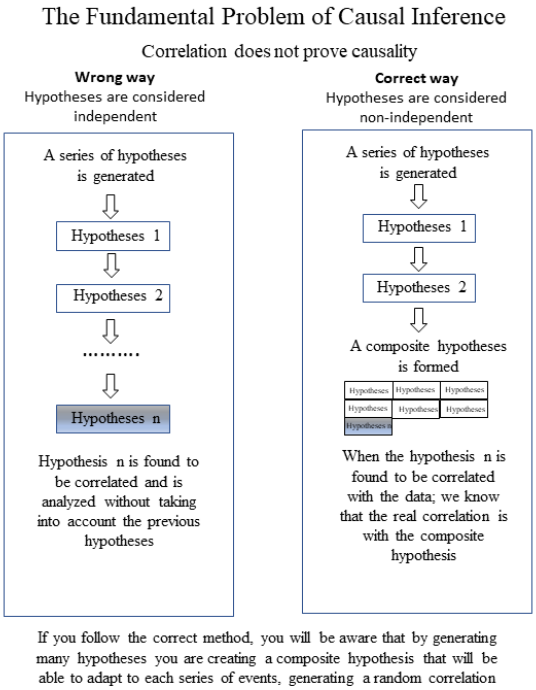


*Figure 1: shows that the correct way is to consider the generated hypotheses as non-independent. In this way, one develops the awareness that by generating many hypotheses one creates a composite hypothesis capable of generating random correlations.*

## Calculating the probability that the correlation is random

Correctly calculating the probability of getting an equal or better result randomly involves changing our approach to statistics. The approach commonly used in statistics is to consider the data produced by one method independent of the data produced by different methods. This way of proceeding seems the only possible one but, as we will show in the following paradox, it leads to an illogical result, which is instead solved by considering the data as non-independent.

We think to have a computer with enormous computational capacity that is used to develop hypotheses about a phenomenon that we want to study. The computer works as follows: it creates a random hypothesis and then performs a statistical test. At this point, we ask ourselves the following question: can there be a useful statistical test to evaluate the results of the hypothesis generated?

If we answer yes, we get an illogical result because our computer would always be able, by generating a large number of random hypotheses, to find a hypothesis that passes the statistical test. In this way, we arrive at the absurd conclusion that it is possible to create knowledge randomly, because it is enough to have a very powerful computer and a statistical test to understand every phenomenon.

If we answer no, we get another illogical result because we are saying that no hypothesis can be evaluated. In practice, the results of different hypotheses are all equivalent and indistinguishable.

How can we solve this logical paradox? The only way to answer the question, without obtaining an illogical situation, is to consider the results obtained from different methods depending on each other. A function that meets this condition is the probability of getting an equal or better result at random. Indeed, the calculation of this probability implies the random simulation of all the actions performed. Hence, random attempts increase the number of actions performed and consequently increase the probability of obtaining an equal or better result randomly.

For this reason, generating random hypotheses is useless, and therefore if you use this parameter, it is

possible to evaluate the data and at the same time it is impossible to create knowledge by generating random hypotheses. Considering the hypothesis as non-independent is a fundamental condition for correctly calculating of the probability that the correlation is random. The probability of getting an equal or better result at random meets this condition.

The dependence of hypothesis on each other has profound implications in statistics, which will be discussed in the next section.

## Consequences of the non-independence of the hypothesis

Consider the tested hypotheses to be dependent on each other when calculating the probability that the correlation is causal leads to three fundamental consequences in statistics.

First fundamental consequence of the non-independence of the hypothesis: our every random action always involves an increase in the probability of a random correlation.

Example: We need to analyze a statistical datum represented by 10 predictions about an event that can only have two results. The 10 predictions are divided into 8 successes and 2 failures. To calculate the probability of obtaining an equal or better result randomly we use the binomial distribution and we get the following value 5.5%. If before making these 10 predictions, we tested a different hypothesis with which we made 10 other predictions divided into 5 successes and 5 failures, the uncertainty of our result changes. Indeed, in this case, we must calculate the probability of obtaining a result with a number of successes greater than or equal to 8 by performing two random attempts consisting of 10 predictions each. In this case, the probability becomes 10.6%, so the fact of having first tested a random hypothesis almost doubled the probability of a random correlation of our second hypothesis. Consequently, increasing the random hypotheses increases the number of predictions that we will have to make, with the true hypothesis, to have a low probability that the correlation is coincidental.

Second fundamental consequence of the non-independence of the hypothesis: every random action of ours and of every other person equivalent to us, always involves an increase of the probability that the correlation is random.

By the equivalent term, we mean a person with the same reputation as us, therefore the data produced by equivalent people are judged with the same weight.

Example: 10 people participate in a project whose goal is the development of an algorithm capable of predicting the outcome of an event that can have only two results. An external person who does not participate in the project but is aware of every attempt made by the participants evaluates the statistical data obtained. All participants make 100 predictions, 9 get a 50% chance of success, one gets a 65% chance of success. The probability that a 65% success is due to a random correlation is obtained by calculating the probability of obtaining a result with a number of successes greater than or equal to 65 by performing ten random attempts consisting of 100 predictions each. The probability obtained, in this way, is 16% instead if he was the only participant in the project the probability would have been 0.18%, therefore about 100 times lower.

Third fundamental consequence of the non-independence of the hypothesis: the calculation the probability that the correlation is random varies according to the information possessed.

Example: 10 people participate in a project whose goal is the development of an algorithm capable of predicting the outcome of an event that can have only two results. In this case, people do not know the other participants and think they are the only ones participating in the project. All participants make 100 predictions, 9 get a 50% chance of success and one gets a 65% chance of success. The participant who obtains a probability of success of 65% independently calculate the probability that the correlation is coincidental. Not knowing that other people are participating in the project, calculate the probability of

obtaining a result with a number of successes greater than or equal to 65 by performing a single random attempt consisting of 100 predictions; the probability obtained is 0.18%. An external person who is aware of every attempt made by the participants calculate the probability that the 65% success rate of one of the participants was due to a random correlation. knowing the number of participants in the project calculates the probability of obtaining a result with a number of successes greater than or equal to 65 by making ten random attempts consisting of 100 predictions each. The probability obtained, in this way, is 16%, a much higher value than the probability calculated by the participant. The probability calculated by the external person using more information is most accurate than the probability calculated by the individual participant. Consequently, the probability obtained by exploiting the greatest number of information must always be considered, in the case of the example, the probability that the 65% success is due to a random correlation is 16%. Therefore, the participant having less information underestimates this probability.

The first and second fundamental highlighting consequence of the non-independence of the hypothesis can be redefined by highlighting the non-randomness of the action.

First fundamental consequence of the non-independence of the hypothesis: our every non-random action always involves a reduction in the probability that the correlation is random.

Second fundamental consequence of the non-independence of the hypothesis: every non-random action of ours and of every other person equivalent to us, always involves a reduction in the probability that the correlation is random.

## How to perform correctly the statistical hypothesis test

About to perform correctly the statistical hypothesis test, It is interesting to note how the non-independence of the hypothesis 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 [3], [4], 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 [5], [6].

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 a big problem: by acting in this way there is the risk of discarding a correct hypotheses and cannot be applied to all fields of research. For example, in finance where the possible inefficiencies of the markets [7], which can be observed, are minimal, adopting very restrictive valuation methods means having to discard almost any hypothesis.

## Conclusion

In this article, we analyzed the fundamental problem of causal inference from a statistical perspective. From this point of view, the problem arises from treating all tested hypotheses as independent of each other. This way of acting is wrong because when we generate a series of hypotheses we are building a composite hypothesis that will tend to adapt and therefore give a random correlation to each of our series of data. **It is believed that this incorrect approach is the cause of the problem of non-reproducibility of scientific results.** Moreover, the increase in computational capacity speeds up hypothesis development, inadvertently creating composite hypotheses that can lead to random correlations.

The probability that a correlation is random is obtained by calculating the probability of obtaining an equal or better result randomly. This calculation can be done, correctly, only by knowing all the hypotheses tested, unfortunately this information is very difficult to have.

For this reason, in modern statistics it is considered fundamental to develop the awareness that each of our compulsive and irrational actions, which leads us to develop and test a large quantity of hypotheses, has as a consequence the generation of random correlations that are difficult to detect.

## Bibliography:


[1] Berdondini, Andrea, “Statistics the Science of Awareness” (August 30, 2021). Available at SSRN: https://ssrn.com/abstract=3914134.

[2] Holland, P. W. (1986) - Statistics and Causal Inference. Journal of the American Statistical Association, 81(396), 945-960.

[3] 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, 53(1), 11–22.

[4] Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (1997). "What if there were no significance tests?" Psychological Methods, 2(4), 315-328.

[5] 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.

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

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



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