The uncertainty of the statistical data

Andrea Berdondini

ABSTRACT: Any result can be generated randomly and any random result is useless. Traditional methods define uncertainty as a measure of the dispersion around the true value and are based on the hypothesis that any divergence from uniformity is the result of a deterministic event. The problem with this approach is that even non-uniform distributions can be generated randomly and the probability of this event rises as the number of hypotheses tested increases. Consequently, there is a risk of considering a random and therefore non-repeatable hypothesis as deterministic. Indeed, it is believed that this way of acting is the cause of the high number of non-reproducible results. Therefore, we believe that the probability of obtaining an equal or better result randomly is the true uncertainty of the statistical data. Because it represents the probability that the data is useful and therefore the validity of any other analysis depends on this parameter.

Introduction

Any result can be generated randomly and any random result is useless. Traditional methods [1] and [2] define uncertainty as a measure of the dispersion around the true value and are based on the hypothesis that any divergence from uniformity is the result of a deterministic event. The problem with this approach is that even non-uniform distributions can be generated randomly and the probability of this event rises as the number of hypotheses tested increases. Consequently, there is a risk of considering a random and therefore non-repeatable hypothesis as deterministic. Indeed, it is believed that this way of acting is the cause of the high number of non-reproducible results [3] and [4]. Therefore, we believe that the probability of obtaining an equal or better result randomly is the true uncertainty of the statistical data, because it represents the probability that the data is useful and therefore the validity of any other analysis depends on this parameter.

In addition, we will also address the problem of determining the correct method of calculating the probability of obtaining an equal or better result randomly. Regarding this topic, we will see that the fundamental point, in calculating this probability value, is to consider the statistical data dependent on all the other data generated by all the tested hypotheses.

In this way, we obtain a ‘paradoxical’ situation, because we can have a series of results where the outcome does not depend on the previous events, but the uncertainty associated with the single result turns out to be dependent on the previous events. So, how can we consider these results as dependent or independent? Two events are independent if there is no kind of dependence between them. In this case, since a dependency has been generated in the calculation of their uncertainty, the events can no longer be considered independent of each other.

Furthermore, as we will see later, the problem of causal inference defined as the inability to understand whether a correlation also implies causation depends precisely on considering individual hypotheses as independent.

Considering the statistical data as non-independent has fundamental implications in statistical analysis. Indeed, all our random actions are not only useless, but will increase the uncertainty of the statistical data. For this reason, in the following article [5], we highlight the importance of acting

consciously in statistics.

Furthermore, the evaluation of the uncertainty of the statistical data will be possible only by knowing all the attempts made. In practice, the calculation of uncertainty is very difficult because not only we must consider all our attempts, but we must also consider the attempts made by every other person who is performing the same task as us. In this way, the uncertainty of our statistical data also depends on the actions performed by the people who are working our own analysis. Indeed, a group of people who belong to a research network all having the same reputation who all work on the same problem can be considered with one person who carries out all the attempts made. Consequently, the calculation of uncertainty becomes something relative that depends on the information we have.

Finally, we will see how this new definition of uncertainty allows us to solve the fundamental problem of causal inference (correlation does not prove causality). **Indeed, the indeterminacy between correlation and causality is not something absolute but derives from the error of considering the hypotheses as independent of each other.** For example, if I test ten hypotheses and the tenth hypothesis tested has a good correlation, this does not mean that the tenth hypothesis is correlated but it means that among ten hypotheses there is one that is well correlated. If instead of testing 10 hypotheses I test 10 thousand whatever my data will be, there will always be a hypothesis, among these 10 thousand, which will be well correlated with my data. In this case, the probability of obtaining an equal or better result randomly, generating 10 thousand random hypotheses, will be very high. Therefore, I will be sure that, in this case, correlation does not imply causation. As a result, statistical data uncertainty defined as the probability of obtaining the same or better result randomly also represents the probability that correlation does not imply causation.

Definition of uncertainty

The aim of the definition of uncertainty of the statistical data that we are going to give is to determine a parameter that is linked to the repeatability of the result and that is universal and therefore, independent of the system in which we perform the statistical analysis.

We define the uncertainty of the statistical data as the probability of obtaining an equal or better result randomly.

This definition considers the statistical data as a forecast, so a forecast is repeatable only if the process that generated it is non-random. Consequently, the calculation of uncertainty involves determining the type of process that generated the result. We can distinguish cognitive processes from random processes by their statistical property of generating non-reproducible results in a random way. Indeed, by using the information on the system, on which we are performing a measurement, we can increase our probability of forecasting and this leads to a consequent decrease in the probability of obtaining the same result randomly.

It is interesting to note that the repeatability of the statistical data and non-randomness of the process that produced it are two equivalent concepts. Indeed, the information leads to the repeatability of the result and at the same time generates results that cannot be reproduced randomly.

To understand the definition given, we report the following example: We have to analyze a statistical datum represented by 1000 predictions on an event that can have only two results. The 1000 predictions are divided into 600 successes and 400 failures. To calculate the probability of obtaining an equal or better result in a random way, we use the binomial distribution and we obtain the following value $1.4 \cdot 10^{-8}\%$.

Now, instead, let us consider a statistical datum represented by 10 predictions divided into 8 successes

and 2 failures. In this case, the probability of getting an equal or better result randomly is 5.5%.

Comparing the two results, we note that in the first case, although the number of successes is only 60%, the uncertainty is almost zero, while in the second case, with a probability of success of 80%, the uncertainty is much higher. This difference is due to the fact that the definition given, as mentioned, concerns only the repeatability of the result and not its accuracy. Therefore, it is a value that decreases as the repetition of the result increases. The definition we have given of uncertainty represents a qualitative and not a quantitative measure of the statistical data. In practice, it tells us if there is a deterministic component in the data collected.

The fundamental point to understand is that the probability that statistical data is completely random and the estimate of its random component (dispersion around the true value) are two parameters that are only partially dependent on each other. The first decreases as the number of repetitions of the measurement increases, the second does not and this is one of the reasons, why the traditional definition of uncertainty, in many cases, is not significant with regard to the repeatability of the result.

The problem, as we have seen in the examples, is that there is always a greater or lesser probability that a purely random process generates the result. In this case, any analysis turns out to be wrong, for this reason, this value is considered the true uncertainty of the statistical result.

Calculation of the uncertainty of the statistical data

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 (random correlation). 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, as a measure of uncertainty, it is possible to evaluate the data and at the same time it is impossible to create knowledge by generating random hypotheses.

Considering the statistical data as non-independent is a fundamental condition for correctly calculating the uncertainty. The probability of getting an equal or better result at random meets this condition.

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

Consequences of the non-independence of the statistical data

Considering the statistical data dependent on each other in the calculation of uncertainty leads to three fundamental consequences in statistics.

First fundamental consequence of the non-independence of the statistical data: our every random action always involves an increase in the uncertainty of the statistical data.

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 uncertainty 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 an acceptable uncertainty.

Second fundamental consequence of the non-independence of the statistical data: every random action of ours and of every other person equivalent to us, always involves an increase in the uncertainty of the statistical data.

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 uncertainty of the static data of the participant who obtains a probability of success of 65% 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 statistical data: the calculation of the uncertainty 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 calculates the uncertainty of the result obtained. 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 calculates the uncertainty of the participant's statistical data, which obtains a probability of success of 65%. It then 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 uncertainty calculated by the participant. The uncertainty value calculated by the external person using more information is most accurate than the uncertainty value calculated by the individual participant. Consequently, the uncertainty value obtained by exploiting the greatest number of information must always be considered, in the case of the example, the most accurate uncertainty is that of 16%.

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

First fundamental consequence of the non-independence of the statistical data: our every non-random action always involves a decrease in the uncertainty of the statistical data.

Second fundamental consequence of the non-independence of the statistical data: every non-random action of ours and of every other person equivalent to us, always involves a decrease in the uncertainty of the statistical data.

Solving the fundamental problem of causal inference

The fundamental problem of causal inference defines the impossibility of associating a link between correlation and causality, in other words: correlation does not prove causality. In this paragraph, we will see how this problem arises from an error and how the uncertainty of the statistical data (the probability of obtaining the same result randomly) represents the probability that correlation does not imply causality.

On the internet, you can find hilarious correlations between very different events, 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 a fundamental piece of information, because if I have a database of a million events, whatever my data is, there will always be a event that will be well correlated with my data.

So, if I generate a million random hypotheses, I will succeed with a near 100% probability of finding a hypothesis that is correlated with my data. So, having about a 100% chance of being able to get the same correlation randomly, I have about a 100% chance that the correlation doesn't also imply causation.

Instead, if we generate a single hypothesis and it is well correlated with the data, in this situation, almost certainly the correlation also implies causality. 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 obtain a good correlation with a single attempt, only if we have knowledge of the process that generates the data. And it is precisely this knowledge that also determines a bond of causality.

Conclusion

The traditional definition of uncertainty implies considering true, for non-homogeneous data dispersions, the hypothesis that the result is not completely random. We consider this assumption the main problem of the definition of uncertainty and the primary cause of the high number of non-reproducible results. Indeed, whatever the statistical data obtained, there is always a possibility that they are completely random and therefore useless.

This error stems from the fact that the definition of uncertainty was developed in an environment where each method had a strong deterministic component. Therefore, calculating the probability of obtaining an equal or better result at random might seem useless. However, when we apply statistics in fields such as finance, where the random component is predominant the traditional approach to uncertainty turns out to be unsuccessful. It fails for the simple reason that the hypothesis on which it is based may not be true. For this reason, we have defined the uncertainty of the statistical data as the probability of obtaining an equal or better result randomly. Since this definition of uncertainty is not linked to any hypothesis, it turns out to be universal. The correct calculation of this probability value implies considering the statistical data dependent on each other. **This assumption, as we have shown through a paradox, makes the definition of uncertainty given consistent with the logical principle that it is not possible to create knowledge randomly.**

The non-independence of the statistical data implies that each action performed influences the calculation of uncertainty. The interesting aspect is that a dependence is also created between actions performed by different people. **Consequently, the calculation of uncertainty depends on the information in our possession, so it becomes something relative that can be determined only with complete knowledge of the information.**

Finally, we have shown how the fundamental problem of statistical inference (correlation does not mean causality) arises from the error of considering the hypotheses tested as independent of each other. This problem is solved by using the probability of obtaining an equal or better result randomly as the uncertainty of the statistical data. Indeed, in this case, uncertainty also defines the probability that the correlation does not imply causality.

Bibliography:

[1] Bich, W., Cox, M. G., and Harris, P. M. Evolution of the "Guide to the Expression of Uncertainty in Measurement". Metrologia, 43(4):S161–S166, 2006.

[2] Grabe, M .,"Measurement Uncertainties in Science and Technology", Springer 2005.

[3] 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.

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

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

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