International Journal of Business Analytics and Intelligence
8 (2) 2020, 01-03
http://publishingindia.com/ijbai/

The Cause of the Effect

Arnab Kumar Laha*

Introduction

In many problems of science and social science we encounter situations in which we have to determine whether an intervention has been effective or not. In other words we want to ascertain whether the observed effect is really due to the intervention or has just happened by chance. For example, suppose a new diet, say Diet X, is proposed by its manufacturers as an effective way for losing weight for obese individuals. How does one check whether the claim of the manufacturer is correct or not?

Consider an individual who has agreed to adopt Diet X. We measure his weight just before he starts on the Diet X and find his weight to be W_B. Then after a month on Diet X we measure his weight again and find that his weight is W_A. Thus, $D = W_B - W_A$ is the change in his weight during this one-month period when he was on Diet X. Note that $D > 0$ indicates a loss of weight whereas $D < 0$ indicates a gain in weight. Suppose we observe $D > 0$. Can we ascribe this loss of weight to Diet X? If we do not think carefully, the answer is apparently 'Yes'. However, on a little reflection, we would recognise that the loss of weight could have been due to various other reasons such as an illness or a disease which is unknown to the individual. Unless such possibilities are ruled out, it would be incorrect to ascribe the loss of weight of this individual to his adoption of Diet X.

Can we do better? One approach could be that we look at the change in the person's weight when he is on Diet X, say D_X and also when he is on normal Diet, say D_N, and compare D_X and D_N. If $D_X > D_N$, then we can say that the loss of weight is due to Diet X. However, we face an insurmountable problem here- the individual at a time can either be on Diet X or on normal diet. He cannot be simultaneously on Diet X as well as on normal diet. Thus, if a person is on Diet X we can find D_X but it is impossible for us to know what would have happened had he been on normal diet i.e. we cannot obtain D_N. Similarly, for another person who is on normal diet we can obtain D_N but it is impossible for us to tell what would have happened had he been on Diet X i.e. we cannot obtain D_X for this person. Thus, the above plan for comparing D_X and D_N cannot be executed in the absence of simultaneous measurements of D_X and D_N.

You may have realised by now establishing causality is not an easy endeavour except in the simplest of circumstances. This stems from the fact that the same subject, cannot be simultaneously on both the 'treatment' (Diet X) as well on 'control' (normal diet). This allows for the possibility of factors other than the treatment, being responsible for the observed effect. How do we resolve such claims of causality? In this article we briefly discuss a few of these approaches.

Sir Ronald Fisher suggested the use of specially designed experiments for dealing with this problem. Suppose that we have a group of individuals on whom we want to test the effect of Diet X. Specifically, let us assume that all these individuals are men having above average BMI but are otherwise healthy. We want to know the impact of Diet X on this group in terms of reduction of weight after following the Diet X regimen for a month. A possible way to do this is to carry out the following experiment. We randomly allocate the men into two groups - Group X and Group N. All individuals in Group X follow the Diet X regimen for one month and all the individuals in Group N follow their normal diet. Suppose both the groups contain k men each. The weights of the k men in the Group X before the commencement of the Diet X regime is noted and suppose they are $W_{BX,1}, W_{BX,2}, \ldots, W_{BX,k}$. Likewise, the weight of the k men who are on the normal diet group are also noted and let these be $W_{BN,1}, W_{BN,2}, \ldots, W_{BN,k}$. During the month we take actions

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

that ensure that the men in Group X sticks to the Diet X and do not deviate. After the one-month period is over, the weights of all the individuals are again noted. Let, the weights of the individuals in Group X after the one-month period be $W_{AX,1}, W_{AX,2}, \ldots, W_{AX,k}$ and those in Group N be $W_{AN,1}, W_{AN,2}, \ldots, W_{AN,k}$. From these data we compute the change in the weight of all the individuals in Group X in the one-month period as $D_{X,1} = W_{BX,1} - W_{AX,1}, \ldots, D_{X,k} = W_{BX,k} - W_{AX,k}$ and those in Group N as $D_{N,1} = W_{BN,1} - W_{AN,1}, \ldots, D_{N,k} = W_{BN,k} - W_{AN,k}$. Since the allocation of the men taking part in this study to the two groups were done randomly, we expect that no systematic bias would be present that may lead to erroneous conclusions. In general, random allocation ensures that the two groups are homogeneous in all other respects except the treatment. If the average value of the $D_{X,i}$'s () is significantly larger than the average value of $D_{N,i}$'s () we can conclude, with reasonable degree of confidence, that the Diet X regimen is effective in reducing weight in the study group. In other words, we say that Diet X is the cause of the weight loss of individuals in Group X.

Design of Experiments (DoE) is a large field of study that has contributed significantly in improving products and processes in different areas such as agriculture, engineering product and process design, drug development, healthcare process improvement and many more. DoE is often used in the 'Improve' step of DMAIC in Six Sigma for identifying the best alternative among many potential alternatives. The Japanese quality guru, Genechi Taguchi used DoE for product quality improvement. His ideas, known as Robust Quality or Taguchi Methods, focus on using DoE for designing products that can withstand wide variations in field operating conditions.

One of the key questions raised during the ongoing COVID-19 pandemic has been the question of efficacy of the different vaccines proposed by the manufacturers. This raises the question: How does one know whether a vaccine is effective in preventing a COVID-19 infection? It is not ethical to expose vaccinated individuals to COVID-19 infection intentionally. If the vaccine doesn't work then the exposed person can become sick and may even die due to the infection. Thus a different method is required for dealing with this situation. In what is commonly known as the Phase-III clinical trial, a large group of individuals who meet the pre-specified health conditions that make the vaccine safe for administration is identified. All these individuals are given an injection which can either be the vaccine or a placebo (such as distilled water). Half the supplied injections are of the vaccine while the remaining half contains the placebo. The injections look identical in all other respects so that neither the doctor nor the patient can figure out whether s/he received the vaccine or the placebo. (Such clinical trials are called double blind trials). All the individuals are tracked for a pre-specified period of time (say, three months) and all the instances of COVID-19 infection during this period are recorded. An analysis is done comparing the number of cases of COVID-19 in the vaccine group and in the placebo group. If the number of cases in the vaccine group is significantly less than that of the placebo group we conclude that the vaccine is effective in preventing COVID-19 infection.

You may be wondering about the need for administering the placebo. This is important since it might happen that after administration of the injection the behaviour of some of the individuals may change. Some of them may feel that they have received the vaccine and are now immune to the infection. This can make them lax about maintaining the norms of safe behaviour such as wearing of mask, social distancing and frequent hand washing. As a result, they may get exposed to the infection more frequently than others. Since with more exposure the chance of catching the infection increases, this makes the vaccine appear less effective than actual. With a control group that has been administered the placebo it would be possible to identify the impact of such behaviour and a correct understanding of the efficacy of the vaccine can be obtained.

Alternatively, it may happen that due to some reason such as development of herd immunity or the weakening of the infectivity of the virus the chance of catching the infection reduces in the general population. Then most of the vaccinated persons would not catch the infection making the effectiveness of the vaccine look better than actual. Here also the presence of a control group is critical as it would allow for estimation of the impact of these changes which would lead to accurate understanding of the vaccine efficacy.

In some instances it may not be possible to have a random allocation of individuals into different groups. Say for example, we are interested to know the impact of smoking on certain health parameters such as occurrence of mini brain-stroke. It is not possible to ask individuals

to smoke a certain number of cigarettes every day for a prolonged period of time knowing fully well that it may have detrimental effects, such as occurrence of cancer in different parts of the body. So how does one, figure out the impact of smoking on the occurrence of mini brain-stroke? One way to proceed is to create a treatment group by choosing individuals meeting the inclusion criterion (such as smoking a certain number of cigarettes per day but having no history of hypertension, diabetes or heart disease) from the group of all the individuals who are smokers. Similarly, a control group can be created by "judiciously" choosing individuals from the group of all non-smokers. These individuals then need to be tracked for a pre-defined period, say five years and incidents of occurrence of mini brain-strokes are recorded. At the end-of the study period the number of mini brain-stroke events in the two groups are compared to check if the number of such events in the smoking group is significantly larger than in the control group.

However, such observational studies, can encounter many difficulties as events such as mini brain-stroke can happen due to many reasons such as age, daily calorie consumption, stress level etc. Such variables are often called confounding variables. To avoid the problems created by the presence of confounding variables, while creating the control group we attempt to match the attributes of the individuals in this group with the attributes of the individuals in the treatment group, to the extent possible. Without careful matching of individuals in the control group and treatment group we would not be able to draw valid conclusions.

There are many ways of matching individuals for creating the control group. A simple way of matching participants, is to match them on their attributes. We can define a distance between any two individuals, based on the attributes of importance for the researchers. If all the features under consideration are numerical then the Euclidean distance can be used. For example, let A be a smoker in the treatment group whose age, daily calorie consumption and stress level values are x_1, x_2 and x_3 respectively. Suppose we have a group of non-smokers who meet the inclusion criteria and for each of whom we have information about their age, daily calorie consumption and stress level. For the i-th individual let these values be y_{1i}, y_{2i} and y_{3i} respectively. Who among these would be chosen in the control group to match with A? For each individual i, we compute the Euclidean distance of A with the i-th individual as

$$D_i = \sqrt{(x_1 - y_{1i})^2 + (x_2 - y_{2i})^2 + (x_3 - y_{3i})^2}.$$

The individual with the lowest value of D_i is chosen as the match of the individual A in the control group. In this way we attempt to make the treatment and control groups as similar as possible with respect to the confounding variables. Now, if we compare the results of these two groups and observe a significant difference we can be reasonably confident that the difference is caused by the treatment.

Another widely used approach for matching uses the 'propensity score'. This approach of allocation is useful in situations where one or more variables impact the chance of an individual being included in the treatment group. For e.g. it may be the case that higher proportion of individuals with high stress levels are smokers compared to individuals with lower stress levels. In such cases a choice model such as a logistic regression model, is used to estimate the probability of a person with the observed characteristics being in the treated group. Now, consider an individual in the treatment group with propensity score P_i. Among all the probable individuals who are not in the treatment group, the individual with the propensity score closest to P_i is chosen to be included in the control group.

The subject of Causal inference has made great progress in the recent times with development of many exciting new techniques. A very readable introduction to this intriguing subject is the book by Pearl and Mackenzie (2018).

Reference

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect.* Allen Lane.