On Counterexamples and Outliers (CDT-51)

Preprint · January 2021

DOI: 10.13140/RG.2.2.26192.12805/5

1 author:

Luciano da F. Costa
University of São Paulo
734 PUBLICATIONS **13,156** CITATIONS

SEE PROFILE

On Counterexamples and Outliers (CDT-51)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

v 1.0: 2nd Jan. 2021
v 1.1: 10 Jan 2021

Abstract

What is the relationship between counterexamples and outliers? How their presence can impact the validity of logical and statistical models? How are these issues related to the central themes of causality and scientific modeling? The current work develops a relatively informal and brief approach to these subjects, with particular emphasis on the issue that, given a statistical relationship such as correlation, the presence of outliers should not be necessarily taken as invalidating otherwise well-defined statistical relationship.

"Inside the egg, an echidna?"

LdFC.

1 Introduction

Science and technology are, to a great extent, the direct consequence of the human tendency to recognize and model patterns (static and dynamic, e.g. [1]). Simply speaking, models are quantitative, abstract constructs involving mathematical and logical relationships between variables and parameters aimed at understanding real-world phenomenon.

Modeling has been critically important for humans (and also other living beings) as a means not only of better comprehending their environment, but also of *predicting* respective events, an ability with immense importance for survival and perpetuation. However, predicting typically requires a deep understanding of all effects governing a given phenomenon, with *causality*, or *causation*, (not to be confounded with *correlation*, but potentially related to it) being of particular relevance.

Though representing an intuitive concept – most people have a good idea of what causality is, it is not so easily formally defined (e.g. [2]), e.g. in a statistical sense, despite many insightful attempts, constituting a challenging problem extending to the philosophical realm. For the purposes of the present work, we shall assume that, given two subsequent (along time) events A and B, the latter will be said to be *caused* by the former in case B could not have taken place if A had not happened before (or simultaneously). It should be observed that the current work does not constitute a formal or strict ap-

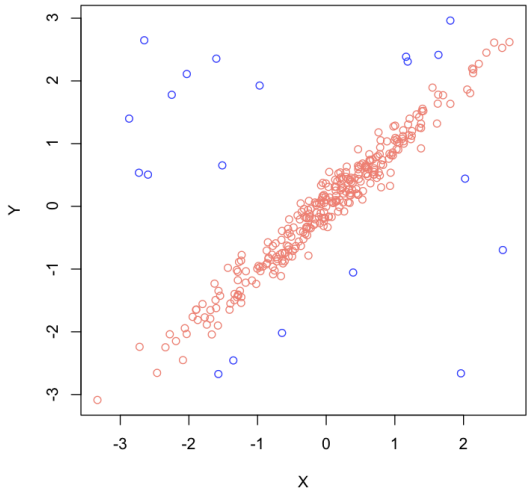

Figure 1: Two measurements (random variables) X and Y of a hypothetical set of entities. The existence of *outliers* (in blue) does not necessarily invalidate the strong tendency (in salmon) that the two measurements X and Y have to vary together (covariance): in statistics, *outliers* do not work as *counterexamples*. Instead, they typically originate from noise or incompleteness (the relationship involves more than two variables). At the same time, covariance should necessarily be taken only a *suggestion* of possible causal relationship between the two variables X and Y, and may be caused by a third, possibly hidden variable. (see Section 4).

proach to causality. Instead, the relatively informal and personal approach this concept provides a particularly relevant context while discussing the possible implications of counterexamples and outliers in logic and statistics.

Going back to the practical aspect of applying models (equations and logical rules) to the real world, the ideal situation would be that every model would take into account all causal respective aspects, paving the way to fully precise predictions. However, as it will be argued in Section 3 (see also, e.g. [1]), models are unavoidably incomplete, so that their predictions can never be fully accurate, implying also that models of the real-world cannot be formally proved. In the face of this result, should we still strive to incorporate as many causal elements as possible into models? The brief answer is, *yes*, because the more causal components are incorporated into a model, the more complete it will be, and more accurate predictions could, in principle, be obtained.

Some of the important consequences of full incorporation of causality on modeling is that *logic* could then be applied to proving or disproving models in the same way as it is typically applied in mathematics, where results could be proved. However, it is necessary to keep in ming the criticality of axions and limitations implied by Gödel's theorem), as well as that nature may not be deterministic.

Because models are unavoidably incomplete, implying eventual error in the respective predictions, it becomes important to resource to probability and statistical regarding the several involved aspects including the ability to estimate the measurement errors and to assign linkelihood to the respective predictions. The very definition of a probabilistic event consists of lack of exact predictability (e.g. [3, 4]). Statistics and probability were developed precisely in order to handle our inability to fully model real world situations.

With a few exceptions, until more recently (e.g. [2]), probability and statistic concepts and methods had seldom contemplated causality, relying instead on important alternative, but less strict, concepts as *correlation* and *independence* which, can only *suggest* causality. Indeed, correlation is less strict than statistical independence, which then is less strict than causality. In other words, the objective identification of a strong correlation between two properties of a phenomenon can only *indicate possible* causality, which is still a very important thing on itself. In a sense, though causality can be seen as the main objective of most statistical analysis, the difficulties of it being directly approached motivates the use of important, though less strict, concepts as correlation and independence.

Interestingly, given that statistical modeling required probabilistic representations of the involved measurements, which are necessarily incomplete in the case of real world problems, it is not possible to use logic to confirm or dismiss a modeling result. Observe that we are *not* referring to the logical framework of a statistical model, which can potentially be objectively verified, but to the own logic of the problem as reflecting nature's laws, which is undermined by the inevitable incompleteness of experiments and knowledge and related representations.

It happens sometimes that counterexamples are sometimes necessarily taken as evidence for invalidating a statistical model. For instance, increases of air pressure have been for centuries taken as a reliable indication of rain being unlikely. Statistically, this result is associated to a significant tendency of one of these measurements to vary inversely with the other, i.e. negative correlation. However, the identification of a *few* days in which rain felt when air pressure was high should not be taken as a disproof of the above mentioned probabilistic model. The key issue here is that these less likely events correspond to *outliers*, being therefore less frequent than those observations following an otherwise well-defined statistical trend, hence our emphasis of the word *few*. In other words, *outliers* do not necessarily invalidate statistical models, though they can provide valuable indication of their incompleteness or limitations.

In brief, we have the quite interesting situation that as the identification of outliers should not necessarily means the demise of a statistical model, neither should the observation of a strong correlation between measurements be taken as a definite proof of causality. Therefore, great attention is required when dealing with the concepts of counterexamples and outliers, as they are intrinsically related to the respectively distinct areas of formal logic and statistics. These questions, as well as related issues, constitute the main motivation and subject of this work.

The present work aims at addressing in a brief and hopefully accessible manner the above important issue related to the concepts of counterexamples and outliers. This endeavor will be made more complete and interesting if contemplated in the light of the important concept of *causality* which, though not often realized, consists the ultimate reason for studying interrelationship between effects, as in the case of correlation.

However, it should be kept in mind that the main objective here, by no means, is not to define or consider causality in a formal manner, which is approached in a more intuitive way contemplating parallels with the flow of causation in logical proofs. This is *not* a work aimed at addressing causality in a complete or formal, strict manner. Instead, our main objective is to discuss the different implications of counterexamples and outliers into the validity (or not) of logic and statistic models, in the light of an informal causality perspective.

We start by presenting the use of *counterexamples* in

logic, which provides a reference for causality. Here, except for the choice of axioms and limitations such as those implied by Göedel's theorem, all representations and operations can take place in a completely causal manner. Informally, it would be as logical propositions and systems could be understood as relating to the flow of truth, especially as implied by the implication relationship. It is this property that allows logic propositions and systems to be proved or not.

In particular, we emphasize the importance of the identity logical relationship for causality, also discussing its relationship with logical implication and sufficient and necessary conditions. The important subject of scientific modeling is then briefly approached, and we argue that, though modeling should be ideally logical constructs, incompleteness stemming from several practical factors imply scientific models to become inherently probabilistic, and therefore not guaranteed to provide infallible predictions or becoming passible of being formally proven. In particular, in modeling causality needs to be replaced by less strict relationships between effects, such as provided by the concept of correlation and independence, which are also briefly reviewed. We then discuss how the probabilistic nature of these concepts, allied to uncertainties and incompleteness of data, as well as inconsistencies in the model, can give rise to *outliers*. This allows us then to address the key issue in the present work, namely that counterexamples are not identical to outliers and therefore do not necessarily imply invalidation of otherwise well-defined statistical relationships suggesting possible causal interactions.

2 Logic and Counterexamples

Informally speaking, logic involves propositions, which can be shown to be true or false, and relationships between then. Arguably, one of the most important logical relationships is that of *identity* between two entities (e.g. variables, propositions, functions, etc.) A and B, which can be expressed as:

$$A \equiv B \tag{1}$$

Given that A and B refer to the same thing, whenever happens to A will be identically reflected by B, and vice-versa. So, the equivalence relationship is intrinsically *causal* in a bidirectional sense, with the additional characteristic of the propagation of this effect being *immediate*, not requiring any temporal lag. In a sense, identity may be understood as a kind of 'instantaneous causality.' Observe that this understanding is not often adopted in causality theories, requiring $t_{future} > t_{past}$ (and not$t_{future} \geq t_{past}$) , but is adopted here for the sake of relating identity to two-directional implication.

Indeed, another particularly important logical concept regards the *implication* of a proposition A into another proposition B, i.e.:

$$A \Rightarrow B \tag{2}$$

which is closely related (if not being identical) to causality.

The logical table for this relationship is presented as follows, where 0 stands for *false* (F) and 1 means *true* (T).

A	B	A ⇒ B
0	0	1
0	1	1
1	0	0
1	1	1

Observe that a false condition A does not necessarily invalidate the verification of B, hence the relationship $0 \Rightarrow 1$ holding true. The own definition of implication makes it plain that the emphasis here is to ensure the propagation of truth (here associated to causality), so that true cannot lead to false, while false has no effect on the conclusion.

Henceforth, we will assume causality to be related to the propagation of the fulfillment of specific verified conditions, with the logical 1 therefore being associated to the observation of a cause, therefore corresponding to flows of logical causality. In this sense, the implication operation means that the verification of a cause A will necessarily lead to the verification of B. Thus, the implication can be understood as the reference logical operation for propagating causality in a directional manner.

Bidirectional causation can be associated to the logical operation of equivalence between two propositions A and B:

$$A \Leftrightarrow B \tag{3}$$

A	B	A ⇔ B
0	0	1
0	1	0
1	0	0
1	1	1

Observe that the equivalence relationship between two propositions can be understood as a particular case of the identity relationship between two entities of any type, as discussed in the beginning of this section.

It is possible to understand causality as being transferred sequentially as a flow along a logic chain in which one proposition successively causes another, such as:

$$A \Rightarrow B \Rightarrow C \ldots \Rightarrow Z \tag{4}$$

which implies $A \Rightarrow Z$, i.e. implication is transitive.

The implication operation provides the basis for understanding the two important concepts of *sufficient* and *necessary*.

More specifically, a *sufficient* condition A to cause B can be logically expressed as:

$$A \Rightarrow B \tag{5}$$

While a *necessary* condition A to cause B can be expressed as:

$$A \Leftarrow B \tag{6}$$

Consequently, a condition A that is both sufficient and necessary to cause B can be expressed as the biconditional implication:

$$A \Leftrightarrow B \tag{7}$$

Another important concept to be used in our discussion of the limitations of modeling concerns the logical operations of *and* and *or* between two propositions A and B, which can be expressed as:

$$A \wedge B \tag{8}$$

$$A \vee B \tag{9}$$

These two logical operations are specified by the following respective logical tables:

A	B	A ∧ B
0	0	0
0	1	0
1	0	0
1	1	1

A	B	A ∨ B
0	0	0
0	1	1
1	0	1
1	1	1

Being associative, these operation may involve several propositions, e.g. $A \wedge B \wedge C$, $A \vee B \vee C \vee D$, or $(A \vee B) \wedge (C \vee D)$. Such combinations provide the natural logical means to represent a combination of hypotheses for an implication, e.g.:

$$(A \wedge B) \vee C \Rightarrow D \tag{10}$$

logic can be applied to other entities, such as variables, functions, etc. This typically takes place by using *logical quantifiers* such as *there exists* ($\exists$) and *for all* ($\forall$).

For instance, we can write:

$$\forall x \in \mathcal{R} \Rightarrow x + 0 = x \tag{11}$$

meaning that for *any* real value x, we have that it is identical to its sum with zero. On the other hand, we write:

$$\exists x \in \mathcal{R} \Rightarrow x + 1 = 3 \tag{12}$$

stating that there is at least one real value $x \in \mathcal{R}$ that when added to 1 yields 2 (in this particular case, this is verified only for $x = 2$).

Though there are additional logic concepts, those briefly discussed above are enough for understanding that the proof of a theorem consists of a chaining of several propositions through which one can go from a logical proposition A to a proposition B so that $A \Leftrightarrow B$, implying that B is caused by A and A is caused by B, being therefore logically equivalent or identical. This construction may also be used, as in Section 3, as a reference for discussing scientific modeling.

We have now reached the point where the critical concept of *counterexample* can be more completely addressed. Let's start with the proposition:

$$\forall x \in \mathcal{R} \Rightarrow x + 1 = 3 \tag{13}$$

This proposition is evidently false, but can this be formally proved so? One particularly simple and effective manner, provided one can find it, is to provide a *counterexample*.

Given that the hypothesis of the above proposition is the *for all* quantifier, it is sufficient to find just one counterexample. In the case of this particular example, we could have, for instance $x = 0$, which yields $0 + 1 = 3$, which is obviously false. Since truth cannot imply a false result, we have formally proven that the above proposition is false.

Contrariwise, if the overall proposition involved the *there exists* quantifier, one manner to prove its validity would be to show that the opposite implication would result for all possible instances of the hypothesis. For example, consider:

$$\exists x \in \mathcal{R} \Rightarrow x = \sqrt{-1} \tag{14}$$

This proposition can be showed not to hold provided it is false for all possible value $x \in \mathcal{R}$.

As already observed, causation may be understood as being associated with the flow of *truth* along a logical or hierarchical (trees) chaining of propositions. Then, as illustrated in Figure 2, once any of the causal links is broken, all subsequent results no longer hold necessarily true (or false).

Interestingly, provided the hyopthesis is verified, mathematical theorems can effectively be understood as flows of truth, or causality. That is why, logic and mathematics provide natural references for better understanding causality in the real world, except for the inevitable incompleteness of any practical approach. It should be also kept in mind that even these abstract approaches are not completely guaranteed because axioms (e.g. the Russell-Zermelo paradox) may eventually found not to be valid, not to mention Goëdel's theorem implications on more systematic formal logic and mathematical systems and the possible fact that nature is intrinsically probabilistic.

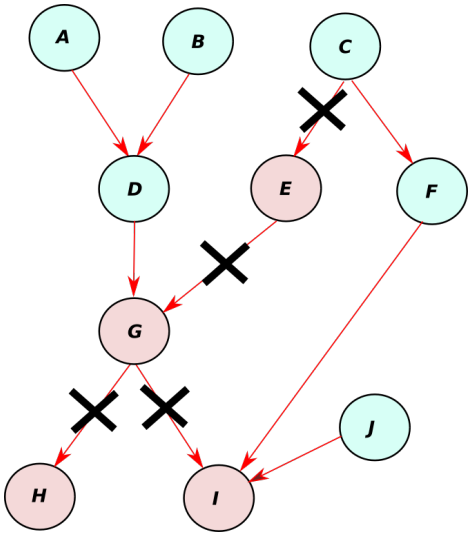

Figure 2: An example of causal propagation of truth, through subsequent logical implications (two-directional implications – i.e. equivalences – may also be involved) between logical propositions (round nodes). One any of the causal links is found to be false (indicated by the cross), all subsequently implied propositions are no longer warranted to be true.

We conclude this section with the observation that:

A single counterexample is enough to disprove a logical proposition.

3 Modeling and Completeness

Being an ability intrinsic to living beings, and particularly to humans (e.g. [1]), the activity of modeling provides exceedingly important resources for *representing*, *better understanding*, and *predicting* real world phenomena, so that suitable decisions can be taken regarding specific objectives. It was thanks to the concept of modeling, supported by the scientific method, that most of the amazing advances in science and technology that continuously reshape modernity were achieved.

Given a phenomenon of interest, some relevant properties (e.g. logical or numerical) have to be identified, in terms of respective *variables* x_i, $i = 1, 2, \ldots, N$ — corresponding to properties allowed to vary as the phenomenon unfolds, as well as *parameters* p_j, $j = 1, 2, \ldots, M$ — respective to properties allowed to change values from one experiment to another, but which are kept constant during any particular experiment realization.

The *hypothesis* of a model can then be logically expressed as a set of propositions:

$$P : \{P_1 \wedge P_2 \wedge \ldots \wedge P_{N_P}\} \tag{15}$$

Provided these hypothesis are verified, the model *result*, or prediction, is typically obtained by the application of a set of logical and numerical equations combining the involved variables and parameters. Here, we will represent these relationships as the set:

$$R : \{R_1 \wedge R_2 \wedge \ldots \wedge R_{N_R}\} \tag{16}$$

Thus, a model can be understood as the logical construct:

$$P \Rightarrow R \tag{17}$$

or, more specifically:

$$P : \{P_1 \wedge P_2 \wedge \ldots \wedge P_{N_P}\} \Rightarrow R : \{R_1 \wedge R_2 \wedge \ldots \wedge R_{N_R}\}$$

This means that *ideally*, given that the hypothesis P hold, it necessary would follow that R is valid. In other words, we have that P is a logically sufficient condition for R.

Let's consider a toy example in which one wants to model the force F that a spring with constant k exerts on a given object when displaced by x from its resting position. The hypothesis is the knowledge of the spring constant as well as the measurement of x, both assumed to be know with absolutely full accuracy. Thus, we may write $P : \{k \wedge x\}$.

Now, we make the assumption that the displacement x is directly proportional to F/k, implying that $F = k\,x$, therefore obtaining the respective model:

$$P : \{k \wedge x\} \Rightarrow R : \{F = kx\}$$

A particular aspect directly associated to causality in modeling is the time interval, delay or lag, that takes for one event A before it can influence, in the sense of causing, another event B.

Though we have briefly conjectured that the identity between two variables intrinsically corresponds to *immediate* causality, most causal effects in the real-world involve some lag. For instance, let's consider the pressing of a piano key. Once the movement is started, it will take a few milliseconds until the hammer hits the string (about 25 ms for *piano* intensity, and 160 ms for *forte* [5]). This is certainly a causal effect, and one involving a delay or time lag. On the other hand, there are situations implying longer delays, such as nymph cicadas metamorphosing into cicadas, which can take 13 or 17 years for some species. Causality delays can be substantially extended along chains of events, so that the overall propagation of the effects may become slower and slower. The existence of long delays in several causations probably accounts for one of the main reasons why we are not more aware of causality, and also why less strict relationships such as correlation are so often employed as resources for better understanding the real-world.

Going back to modeling real world phenomena, something quite interesting happens, namely that *it is impossible to obtain a fully complete set of hypothesis while modeling a real world phenomenon.* First, we have that every measurement is potentially affected by some noise/error related to the accuracy, reliability and resolution of the measuring devices. At the same time, no real world phenomenon can be guaranteed to be fully isolated from the remainder of nature. In addition, it is impossible to incorporate all effects potentially influencing the phenomenon of interest into a given model.

For instance, several physical effects such as electromagnetic field and gravity, take place in terms of asymptotic fields, which never reach completely null values even at very long distances. Thus, it is completely impossible to isolate the falling body in our example from the gravitational influence of every other portion of mass in the universe. Though it is true that these effects may be very small, such influences will nevertheless manifest themselves along a long enough period of time, therefore limiting the time window of the predictions. The influence of slight perturbations become potentially much larger in the case of non-linear phenomena. It is also interesting to observe that error and noise can be understood as specific manifestations of incompleteness.

In the particular case of the spring modeling considered above, the uncertainties in predicting the force as a consequence of incompleteness of the approach stems from the non-exhaustive facts: (i) k and x cannot be determined with full accuracy as a consequence of measurement error and noise; (ii) the unavoidable presence of other objects (masses) in the environment of the considered system exert gravitational effects on the mass of the spring; (iii) it is impossible to have a spring with fully uniform mass distribution; (iv) the spring may deform or break during the measurements; and (v) the hypothesis that $x \propto F/k$ may not be observed, e.g. as a consequence of (iii) or (iv).

In practice, given all the above observed limitations caused by incompleteness, not mentioning the fact that nature could be inherently probabilistic, scientific models can never be fully accurate, and *as such cannot be completely proved or disproved* by formal logical means. That is why it is said that scientific results are true as long as no systematic inconsistencies are observed. That is also the reason why we cannot prove theorems related to the real-world, and why the successive integration of scientific models can neither be verified in a definitive manner. In other words, when scientific models are integrated, e.g. in a chaining or tree manner, the interconnection of models establishes a network of possibly provisory implications. Were scientific models complete, and causal, scientific advancement could proceed through logical means analogous to the flow of causality in propositional proofing.

In a more informal perspective, we could also understand models as corresponding to *rules*, which popular wisdom acknowledges always to have *exceptions*. Observe that, in this context, the exceptions (relatively rare, as implied by the own name) are properly *not* generally taken as invalidating the rule.

Because the representations associated to models can never be complete, the identification of these inconsistencies to a level enough to make the model dubious need to incorporate concepts from probability and statistics, which are addressed in the following section.

Be that as it may, we should also keep in mind that, despite the modeling limitations, the respectively associated scientific method *remains our best and most accurate manner to understand and predict real world phenomena.*

4 Correlation and Outliers

Probability and statistics were developed as powerful means for addressing measurements and phenomena about which we cannot be completely certain (which turns out to be virtually every real world phenomenon). Actually, even the certain event is commonly understood, for completeness' sake, as being a random phenomenon.

Real world measurements can therefore be understood as corresponding to respective *random variables*, which are conveniently described (actually, modeled) by respective *probability density functions* (e.g. [3, 4, 6, 7]). For instance, the outcome of a completely unbiased dice (an abstract construct not feasible in the real world) can be expressed as X, associated to the probability density function $p(x) = 1/6(\delta(1)+\delta(2)+\ldots+\delta(6))$, where $\delta(n)$ is the Dirac delta function placed at n.

Probability density functions are particularly useful because they allow us to make sensible predictions about specific events. More specifically, we have that the probability of observing a value X in the interval $[a, b]$ can be exactly obtained as:

$$P(a \leq X < b) = \int_a^b p(x)dx \tag{18}$$

While random variables are fully characterized by their respective probability density functions (or, alternatively, by respective statistical moments), joint relationships between pairs of random variables are typically addressed in terms of *joint statistical moments* such as *correlation* (e.g. [7])), which is formally defined as:

$$\begin{aligned} K[X,Y] = E\left[(X - E[X])\,(Y - E[Y])\right] = \\ = E[X\,Y] - E[X]E[Y] \end{aligned} \tag{19}$$

where $E[X]$ is the *expectance* (associated to the sample mean or average) of the random variable X, being

respectively defined as:

$$E[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx \tag{20}$$

It can be verified that $K[X,Y]$ will tend to be positive whenever X tends to vary jointly with Y (i.e. they both tend to increase or decrease together). Contrariwise, $K[X,Y]$ will tend to be negative when the two variables undergo opposite tendency of variation. When X presents neither of these two types of joint variation, the covariance $K[X,Y]$ becomes zero, and the two random variables are said to be *uncorrelated*.

The tendency of two random variables to vary together can alternatively be expressed, in a more normalized manner, in terms of the respective *Pearson correlation coefficient* (e.g. [7]), which is given as:

$$\mathcal{R}[X,Y] = E\left[\frac{(X - E[X])}{K[X,X]}\,\frac{(Y - E[Y])}{K[Y,Y]}\right] \tag{21}$$

The random variable $\tilde{X} = \frac{(X-E[X])}{K[X,X]}$ corresponds to the *standardization* of the original counterpart X.

Observe also that $-1 \leq \mathcal{R}[X,Y] \leq 1$, therefore ranging from full anti-correlation to full correlation, with $\mathcal{R}[X,Y] = 0$ indicating complete uncorrelation.

Another interesting statistical concept associated with the possible relationship between two random variables concerns *statistical independence*. More formally speaking, two random variables X and Y, described by respective probability density functions $p(x)$ and $p(y)$ are said to be *independent* if and only if their respective *joint probability density function* can be expressed as $p(x,y) = p(x)\,p(y)$, or, in other words, the function $p(x,y)$ is not *separable*.

It follows necessarily from the fact that two random variables X and Y, with respective joint probability density function $p(x,y)$, to be independent that:

$$E[X\,Y] = E[X]\,E[Y] \tag{22}$$

Then, from Equation 21, we have that:

$$K[X,Y] = E[X\,Y] - E[X]\,E[Y] = 0 \tag{23}$$

meaning that the two statistically independent variables X and Y are also necessarily uncorrelated. Though statistical independence implies uncorrelation, the latter does not necessarily implies independence. In this sense, independence is a more strict property not share by every case of uncorrelated variables. Figure 3 illustrates the joint probability density function of two variables X and Y that are uncorrelated but not independent.

As the Pearson correlation coefficient provides an objective indication of the joint tendency of two random variables to vary together, the key question now regards:

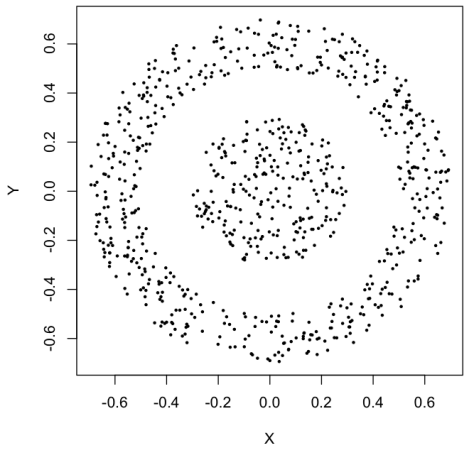

Figure 3: Scatterplot respective to an example of joint probability density function $p(x,y)$ with X and Y being independent but not correlated. As demanded by statistical independence, it is impossible to separate $p(x,y)$, of which this figure illustrates a possible sampling, into two univariate probability density functions $p(x)$ and $p(y)$ so that $p(x,y) = p(x)p(y)$.

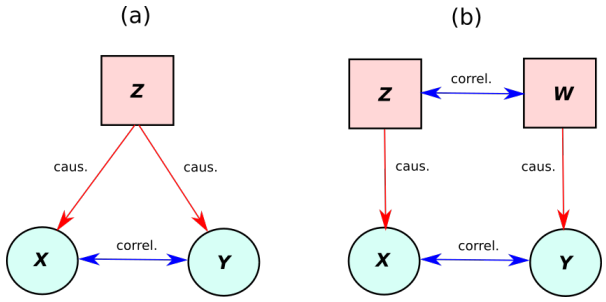

Figure 4: (a) A simple situation in which a random variable Z causes two other variables $X \leftarrow 2Z + noise$ and $Y \leftarrow 3Z + 2 + noise$, resulting in a positive correlation between X and Y. (b) In this case, the random variables X and Y are respectively caused by two correlated random variables Z and W. Observe that, since there is no definitive formal statistical definition of causation, the red arrows in these figures may be understood in the more informal sense that the the removal of X will have no impact whatsoever on Y (X non-causes Y).

how and to which extent this coefficient can be be related to *causality*? Figure 4 illustrates two related situations.

In Figure 4(a), we have two random variables being 'caused' by a common, possible hidden (unknown or unaccessible) random variable Z. For instance, we may write $X \leftarrow 2Z + noise$ and $Y \leftarrow 3Z + 2 + noise$. In case

Since the removal of X can have no implication on Y whatsoever, we have the X and Y are not causally related. However, they are evidently positively correlated.

A similar situation can be inferred from Figure 4(b). Now, we have the two variables $X(t)$ and $Y(t)$ being caused by two respective random variables Z and W that are also correlated. For instance, we may have $X = cos(\omega t)$ and $Y = cos(\omega t + \psi)$.

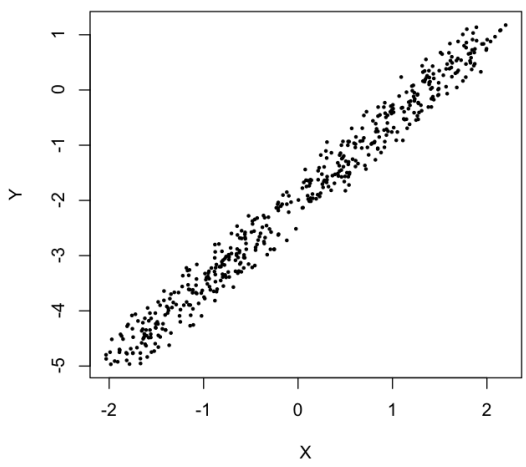

Figure 5: Example of scatterplot between the random variables $X \leftarrow 2Z + noise$ and $Y \leftarrow 3Z + 2 + noise$, where *noise* stands for uniformly random noise in the range $[-0.25, 0.25]$ and Z is also uniformly distributed in $[-1, 1]$. The respectively obtained Pearson correlation coefficient was 0.99.

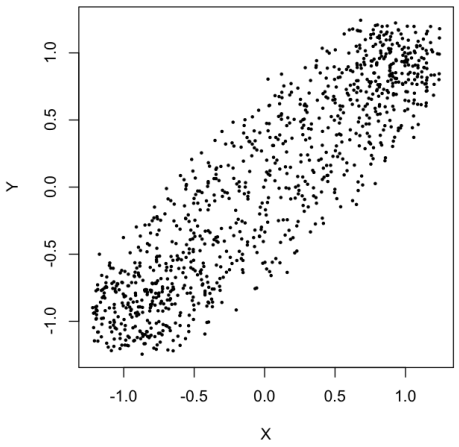

Figure 6: Example of scatterplot between the random variables $X \leftarrow Z + noise$ and $Y \leftarrow W + noise$, where *noise* stands for uniformly random noise in the range $[-0.25, 0.25]$, $Z = cos(\omega t)$ and $W = cos(\omega t + \psi)$, with $\psi = \pi/9$. The respectively obtained Pearson correlation coefficient was 0.90. Observe that the resulting correlation depends on the value of ψ. For instance, null correlation is obtained if $\psi = 0$.

Again, we have an example in which the correlation between X and Y is just a consequence of correlation between Z and W being propagated in a causal way. However, X can by no means cause Y, and vice versa.

A similar situation holds between statistical independence an causality, in the sense that, as with correlation, they do not correspond to the same statistical concept. As such, correlation and independence between two random variables X and Y can only be taken as indications of *possible* causation between those variables. Even with these limitations, these two concepts have played a very important role along the history of science, being to the present day systematic and widely applied to a vast range of problems. As such, they remain important concepts that need to be properly understood.

Given the difficulties in fully understanding how an observed correlation is *not* a consequence of causality, it is instructive to consider the opposite, i.e. how causality may imply (or not) correlation. Insights about this important issue may be readily derived from the above spring model. first, let's suppose that the law $F = k\,x$ is indeed observed, and that k is also known to full precision and that F can also be measured with full accuracy while the estimation of x incorporates a normally distributed error ε with zero means and variance 0.01, so that the observed values can be expressed as $\tilde{x} = x + \varepsilon$. Figure 7 depicts a possibly observed result.

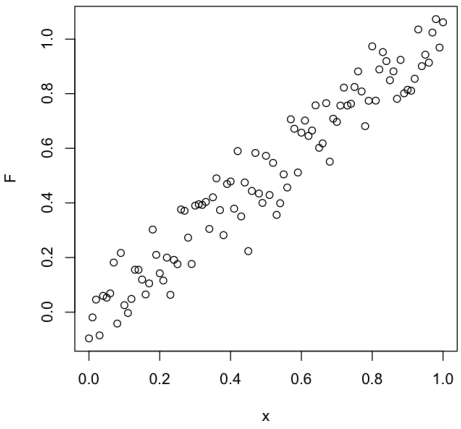

Figure 7: The force F exerted by a spring with respective constant known without error measured with full accuracy in terms of observations of x, affected by error ε normally distributed with null mean and variance 0.01. Instead of the otherwise expected straight line, reflecting the law $F = k\,x$, an elongated distribution of points is observed. The Pearson correlation coefficient for this particular scatterplot is 0.96.

Regarding causation, it is reasonable to assume that the force F is being caused by the displacement x, since the removal of the spring is the same as making $x = 0$, therefore implying no observable force. However, the displacement x may also be understood as being caused by the presence of the force $F \neq 0$. Though a definite conclusion on these causal relationships is hard to be reached, involving even philosophical issues, we will assume that the force F and the displacement x are mutually caused. In case all values could be observed with full precision, a perfectly straight relationship would have been observed in the $F \times x$ graph, suggesting a well-defined causal relationship between the two observed variables.

Now, going back to Figure 7, the presence of uncertainties while measuring x has implied in a substantial dispersion of points that are, nevertheless, substantially

correlated. Therefore, at least for this particular example, we have that a dispersed, but still strongly correlated, distribution of points may still be understood as an indication of possible causality. This evidence can be further investigated not only by trying to make more accurate measurements, but also by devising additional experiments in which the spring is suddenly removed, or the force varied. Incidentally, these planned modifications of the experimental set-up provides one of the main motivations in interesting approaches to causality as described in [2].

Another quite important fact is that a causal relationship of a variable X on Y may not necessarily imply respective correlation. This is typically the case when X is related to Y through a strongly non-linear formula, such as $Y = X^2$ (see Fig. 8). This important fact has motivated alternative measurements of relationships between variables, including mutual information (e.g. [8]), which can be used to quantify relationships more general than linear correlation.

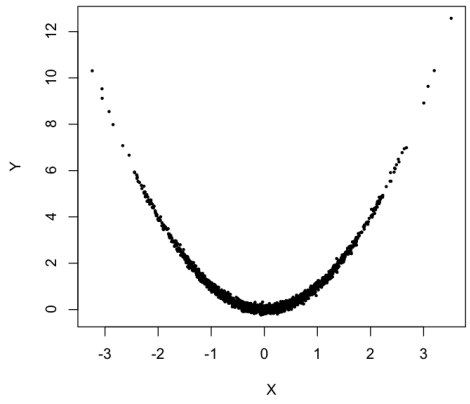

Figure 8: Example of scatterplot obtained from two random variables X ad Y which, though presenting a well-defined non-linear relationship, leads to a very low Pearson correlation coefficient (-0.005). Such relationships can ben more properly characterized by using approaches such as the mutual information between the variables.

In addition to the observation of a disperse relationship between two or more variables as providing uncertainties about the verified relationship, it is also important to consider wether the presence of *outliers* may be eventually taken as a possible evidence against the validity of a possible (causal or non-causal) relationship between two random variables, or measurements, X and Y.

One particularly important issue concerns how the presence of *outliers* can affect the correlation between two random variables (see Figure 1). A first important step is to define in a more formal means what an outline is. Basically speaking, given a scatterplot corresponding to the sampling of a joint probability density function, outliers can be understood as points that are too distant from the main mass of the distribution (e.g. [9]). For instance, we can take as outliers all points in the region of the scatterplot so that the area (volume or hypervolume) of the respective joint probability density in that specific region is smaller than a given threshold $0 < T < 1$.

However, what are the reasons (or cause) for outliers appearing in a given scatterplot? Quite interestingly, the generally correspond to *the same reasons* why a model is incomplete. Therefore, we have that: (i) noise or error in a measurement may cause one or more outliers; (ii) an outlier may be implied by the fact that one or more variables or parameters being left out (intentionally or not) of the model; and (iii) quite distinctly, outliers may be produced as a consequence of the assumed model being intrinsically wrong. Out of these three important causes of outliers, *only in case (iii) their presence should be taken as evidence for invalidation of the model*, and therefore of eventual causal relationship between the observed measurements. At the same time, outliers may provide indication about incompleteness or limitations of the respective model.

It is interesting to observe that the presence of a substantial number of outliers should indeed be taken as serious evidence for questioning a model. However, we have that the own concept of outlier, as discussed above, assumes them to appear with a relatively small probability, so that they should not represent the majority of the observed relationship.

In summary, we have that the presence of dispersion or outliers in an observed (linear or not) relationship between two or more variables should not be necessarily taken as a conclusive indication of non-validity of an otherwise well-defined relationship/model, or even of possible causal interrelationships. However, these indications suggest the measurements to be taken with high accuracy, that additional effects (variables and parameters) may need to be incorporated, or, ultimately, that the model is indeed wrong and that eventual causal relationships are consequently also non-verified, therefore corresponding to artifacts.

In the light of the above discussion, we conclude this section by observing that:

One or more outliers, provided they are not too many, do not necessarily imply the invalidation of an otherwise well-defined statistical relationship.

5 Concluding Remarks

Having brief and informally discussed propositional logic, counterexamples, modeling, completeness, correlation and outliers, we have reacher a better position from which to appreciate the relationship between counterexamples and outliers from the perspective of causality. Indeed, several related verifications have been made, which are summarized as follows:

$$\begin{aligned}
\textbf{\textit{counterexamples}} &\neq \textbf{\textit{outliers}}.\\
\textit{counterexamples} &\Rightarrow \textit{invalidation of a proposition.}\\
\textit{outliers} &\not\Rightarrow \textit{invalidation of correlation.}\\
\textit{outliers} &\not\Rightarrow \textit{invalidation of dependence.}\\
\textit{outliers} &\not\Rightarrow \textit{invalidation of causality.}\\
\textit{correlation} &\not\Rightarrow \textit{dependence.}\\
\textit{correlation} &\not\Rightarrow \textit{causality.}\\
\textit{correlation} &\Rightarrow \textit{possible dependence.}\\
\textit{correlation} &\Rightarrow \textit{possible causality.}\\
\textit{dependence} &\not\Rightarrow \textit{causality.}\\
\textit{dependence} &\Rightarrow \textit{possible causality.}\\
\textit{dependence} &\Rightarrow \textit{uncorrelation.}\\
\textit{causality} &\Rightarrow \textit{possible correlation.}
\end{aligned}$$

It should be kept in mind that the above entries that are related to causality are not completely formal or definitive, reflecting the challenge in defining causality in a completely objective and consensual manner.

Given such a large number of relationships between concepts related to counterexamples and outliers, it should not be too much surprising that those two concepts are sometimes applied in not necessarily the most suitable way.

All in all, counterexamples should not be taken as being identical to outliers, with former being used only to invalidate a proposition, while the latter correspond to a relatively small number of points typically found at the fringe of statistical relationship (see Fig. 1). However, outliers have been sometimes unjustifiably used as necessary counterexamples to otherwise significant statistical relationships which may suggest causal interactions, possibly reflecting a seeming analogy with the role of counterexamples in the proof of logical propositions and theorems and taking outliers as if they corresponded to logical counterexamples.

Several of the concepts addressed in a relatively brief and informal way in the present work constitute subject of important ongoing research by several researchers, in particular regarding causality and its implications, and continuing advances could be expected.

Acknowledgments.

Luciano da F. Costa acknowledges CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2) for financial support.

References

[1] L. da F. Costa. Modeling: The human approach to science. Researchgate, 2019. https://www.researchgate.net/publication/333389500_Modeling_The_Human_Approach_to_Science_CDT-8. [Online; accessed 1-Oct-2020.].

[2] J. Pearl. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2008.

[4] L. da F. Costa. Statistical modeling. https://www.researchgate.net/publication/334726352_Statistical_Modeling_CDT-13, 2019. [Online; accessed 10-Apr-2020.].

[5] A. Askenfelt and E. V. Jansson. From touch to string vibrations. *Pubs. of the Royal Academy of Music*, 64:39–57, 1990.

[6] R. A. Johnson and D.W. Wichern. *Applied multivariate analysis*. Prentice Hall, 2002.

[7] L. da F. Costa. Multivariate statistical modeling. https://www.researchgate.net/publication/340442989_Multivariate_Statistical_Modeling_CDT-26, 2019. [Online; accessed 10-Apr-2020.].

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[9] L. da F. Costa, F. A. Rodrigues, C. C. Hilgetag, and M. Kaiser. Beyond the average: Detecting global singular nodes from local features in complex networks. *EPL (Europhysics Letters)*, 87(1):18008, jul 2009.

Costa's Didactic Texts – CDTs

CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also incorporate new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Each CDT focuses on a limited set of interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provided in order to provide some complementation of the covered subjects.

Observe that CDTs, which come with absolutely no warranty, are non distributable and for non-commercial use only.

Please check for new versions of CDTs, as they can be revised. Also, CDTs can and have been cited, e.g. by including the respective DOI. Please cite this CDT in case you use it, so that it may also be useful to other people. The complete set of CDTs can be found at: https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs, and a respective guide at: https://www.researchgate.net/publication/348193269_A_Guide_to_the_CDTs_CDT-0