Statistical Causality Structure on Prostate Cancer in Puerto Rico and USA

by

Andrés G. Alvarado Alvarado
Abstract

The purpose of this topic is to find specific statistical definitions to connect it to causal conditions. Also, use multiple data-sets of experiments to show how the statistics values are related to causality given the condition. I will use TETRAD Program to find possible causal models for the given data set.
Contents

Acknowledgment and dedication 3

1 Introduction 4

2 Statistical Causality 4
   2.1 Definition ......................................................... 4
   2.2 Bradford Hill’s Causal Criteria .................................. 7

3 Types of Causalities 11
   3.1 Introduction ........................................................ 11
   3.2 Token Causality .................................................... 12
   3.3 Type-level causality .............................................. 14

4 Causal Strength 15
   4.1 Spurious vs Genuine ............................................... 15

5 Method to reduce error in finding causal relationship for Experimental studies using data sets 22

6 Experiments 23
   6.1 Causality in patient’s diagnostics on Glucose level .............. 23
   6.2 Causality in patient’s coronary arteries disease .................. 29
   6.3 Causality in Prostate cancer 1985 study .......................... 35
   6.4 Causality in Prostate Cancer between patients in the USA and PR .... 42
   6.5 Causality of USA and PR Prostate Cancer data using Ang Li’s Statistical analysis to find the significant variables between specific population .................. 48

7 Conclusion 55

8 Appendix: Graphs definition summary 56

9 References 58
List of Figures

1. Correlation between the variables .................................................. 25
2. Causal graph of the experiment ........................................................ 26
3. Possible causal models: the numbers in green in this case are the variable’s mean and the black one are the regression’s coefficient. .................................................. 27
4. Statistical data fitness ........................................................................ 28
5. Correlation between cholesterol variable and Duration of symptoms of CAD .................................................. 29
6. Possible causal graph ........................................................................... 30
7. Optimal causal graph with their p-value (high p-value indicates that the model fits good, so the higher the p-value, the better is the causal graph) .................................................. 31
8. Possible causal graph for 17-35 years old patients .................................. 32
9. Optimal causal graph for 17-35 years old patients with its p-value respectively .................................................. 33
10. Possible causal graph for 51-65 years old patients .................................. 33
11. Optimal causal graph for 51-65 years old patients .................................. 34
12. Optimal causal graph for 65 and older patients ..................................... 34
13. Possible causal graphs: a) patients given placebo; b) patients given 0.2 mg of estrogen .................................................. 38
14. Possible causal graphs: c) patients given 1 mg of estrogen; d) patients given 5 mg of estrogen; .................................................. 39
15. Optimal causal graphs: a) patients given placebo; b) patients given 0.2 mg of estrogen .................................................. 40
16. Optimal causal graphs: c) patients given 1 mg of estrogen; d) patients given 5 mg of estrogen; .................................................. 41
17. Possible causal graphs: a) 40-60 years old case patients; b) 60-70 years old case patients .................................................. 44
18. Possible causal graphs: c) 70-90 years old case patients .................................................. 44
19. Possible causal graphs: a) 40-60 years old control patients; b) 60-70 years old control patients .................................................. 45
20. Possible causal graphs: c) 70-90 years old control patients .................................................. 45
21. Optimal causal graphs: a) 40-60 years old case patients; b) 60-70 years old case patients .................................................. 46
22. Optimal causal graphs: c) 70-90 years old case patients .................................................. 46
23. Optimal causal graphs: a) 40-60 years old control patients; b) 60-70 years old case patients .................................................. 47
24. Optimal causal graphs: c) 70-90 years old control patients .................................................. 47
25. Optimal causal graphs with their respective p-value: a) case patients; b) control patients .................................................. 49
Acknowledgment and dedication

I want to thank God for making the correct path using his knowledge that always guided me through my life. Also, for giving me an excellent family and a beautiful companion who always supported me. I want to dedicate this project to my brother Nelson who motivated me to continue learning a beautiful topic called Mathematics.
1 Introduction

Causality is commonly used to define future events given their prior connection, similar to prediction in statistics but in this case, the future event is always going to happen given that connection, i.e., cause. In our world, we can describe any possible event who somewhat is related to another as a causal connection, but it’s complicated to know which one is the cause and the effect. This difficulty has been striving statistician in years when it comes in experimental and observational data. In statistics, to describe a causal relationship, we have to organize and analyze well the data so that it would be easier to find causal structures. As in this case, I will try to get the causal relation from raw data-sets.

First, I will give the most accurate description of what is a causal structure in terms of statistics specifically in experiments, then I will explain the types of causality and their uses. Also explain the requirements of statistical analysis so we can construct a causal behavior between the interested variables. In addition, find requirements such as we can describe if a distribution has a strong (or weak) causal relationship with another distribution. I will explain why the background knowledge of an experiment is important in statistical analysis to find causal structures. All these tactics is going to be explain in examples to show how it can facilitate us to find causal structure, in other words, restrict the possibilities of causal models in experiments. Finally, I will find (if possible) the causal relationship between the variables of a prostate cancer’s experiment.

2 Statistical Causality

2.1 Definition

Causality has no formal mathematical definition in general, but in statistics, we can describe necessary conditions for it. Before stating the statistical causal definition, I will define what a Manipulation Variable is. Let M be a distribution who’s equally distributed, i.e., all values of M have the same probability, M is a Manipulation Variable if for such a distribution X, M forces X to have some values (Neapolitan 2004). Notice that X will be consider the "cause distribution" since M will be the variable that it’s outside of the system that will manipulate it (and only it). As for the manipulation variable, we have to be careful how to select it because it can lead us to false assumptions and conclusion, so selecting a manipulation variable such as it forces only one distribution (preferably the causal one), it will consider a good manipulation. Now, stating the definition of the manipulation variable, suppose we have X and Y distribution, also suppose that M manipulates X, then X causes Y if for a value of M, X changes the probability of Y (Neapolitan 2004). In other words, being specifically for randomized experiments, denote $E(Y)$ be the mean of Y, and $E(Y|X)$ being the mean of Y given some value of X, then if:
\[
\frac{|E(Y) - E(Y|X)|}{E(Y)} \geq c, \quad \forall x \in X
\]  

(1)

where \(c\) is define by the statistician, then \(X\) is a possible cause of \(Y\). Now, if \(X\) changes the probability of \(Y\), then we can assume that there is a dependency between them, i.e, they’re correlated, so regression can be applied here, later on I will explain in details. The value of \(c\) depends on your goal as a statistician, since even a slightly change in probability of a distribution, we can consider it as a possible cause. However, a logical pattern must exist so that \(X\) cause \(Y\). Assume that \(X\) is a positive cause of \(Y\), then as \(X\) increase his value, \(P(Y|X) - P(Y) > 0\). For example \(^1\), Suppose I’m a person in the experiment and that \(A = \) ”the effect in my body of taking laxative” and \(B = \) ”me going to the bathroom per day”. Suppose that the distribution \(A\) is a uniform distribution with pdf define as:

\[
f(a) = \begin{cases} 
\frac{1}{50} & 10 < a < 50 \\
0 & \text{otherwise} 
\end{cases}
\]

Where \(a\) is the quantity of ml taken of the laxative. Suppose also that the distribution \(B\) has the following possibilities: \(\{1, 2, 3, 4\}\), with probabilities \(P(B = b) = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{12}, \frac{1}{12}\}\) respectively. Now, \(B\) is more often to happen if I take it (could be that \(A\) is independent,i.e, doesn’t affect me), then lets define the joint distribution:

\[
f(a, b) = \begin{cases} 
- e^{-10} \frac{(800m+1199)}{4(e^{10}-1)} + 3b + 5m & , 10 \leq a \leq 50, b = 0, \ldots, 4 \\
    f(b) & b = 1, \ldots, 4 \\
0 & \text{otherwise} 
\end{cases}
\]

Where \(m = -\frac{1199(-7+3e^{10}+3e^{20}+3e^{30})}{2000(-3+e^{10}+e^{20}+e^{30})}\)

Now computing the mean \(E(B)\) and \(E(B|A)\) which are the following:

\[
E(B) = \frac{1}{2} + 2 \times \frac{1}{3} + 3 \times \frac{1}{12} + 4 \times \frac{1}{12} = 1.75
\]

\(^1\)All the example given here will be fictional for convenience, the purpose is to explain the definitions stated
\[ E(B|A) = \sum_{i=1}^{4} b \ast f(b|a) \]

Suppose that my dosage is \( a = 20ml \) so my body manage to absorb at maximum 20 ml, i.e., \( a \leq 20 \), then the expectation of B given \( A \leq 20 \) is:

\[
E(B|A) = \sum_{i=1}^{4} b \ast f(b|a) = \sum_{i=1}^{4} b \ast \frac{P(A \leq 20, B = b)}{P(A \leq 20)} = 3
\]

As you can see, given \( A \) a certain value in his sample space, it changes B in probability (the distribution in this case) as for:

\[
\left| \frac{E(B|A) - E(B)}{E(B)} \right| = \left| \frac{3 - 1.75}{1.75} \right| = 0.70
\]

More over, for all the space of \( A \), denote \( a \) any value of \( A \), then

\[
E(B|A \leq a) = \sum_{i=1}^{4} b \ast f(b|a) = \sum_{i=1}^{4} b \ast \frac{P(A \leq a, B = b)}{P(A \leq a)}
\]

\[
= \sum_{i=1}^{4} b \ast \int_{10}^{a} \frac{e^{a-10}(800m + 1199)}{4(e^{40} - 1)} + 3b + 5m
\]

\[
= \frac{\sum_{i=1}^{4} b \ast \int_{10}^{a} \frac{1}{40}}{\int_{10}^{40} \frac{1}{40}}
\]

\[
= -\frac{100 \left(-4e^{50}(a - 10)(5m + 9) + e^{a}(800m + 1199) + e^{10}(4a(5m + 9) - 1000m - 1559))}{e^{10}(e^{40} - 1)(a - 10)}
\]

Now we want to show that for all the sample space of \( A \), \( A \) changes the mean of \( B \), i.e. \( E(B|A) - E(B) \neq 0 \), for all the values of \( A \), so solving for \( a \) for the given value of \( E(B|A \leq a) \) we have that:

\[
\Rightarrow E(B|A \leq a) - E(B) = 0
\]

\[
\Rightarrow -\frac{100 \left(-4e^{50}(a - 10)(5m + 9) + e^{a}(800m + 1199) + e^{10}(4a(5m + 9) - 1000m - 1559))}{e^{10}(e^{40} - 1)(a - 10)} - 1.75 = 0
\]

We have that:

\[
a = 10
\]
But the value \( a \) cannot be 10 since \( P(A \leq 10) = 0 \) and \( E(B|A \leq 10) = \sum_{i=1}^{4} b_i \frac{P(A \leq 10, B)}{P(A \leq 10)} = \infty \). Thus we can conclude that \( A \) changes the mean of \( B \) for all the sample space of \( A \). However, we cannot say that \( A \) causes \( B \), because we didn’t proof that \( A \) changes the probability of \( B \). We can show it by finding the values of \( A \) such as \( P(B|A) - P(B) = 0 \), Suppose \( P(B|A \leq a) - P(B) = 0 \), and \( P(A \leq a) > 0 \) then there exist a characteristic such that consuming that certain quantity, it tends to process it in a certain way such that it doesn’t affect my body, so that means that there exist an intermediate variable \( A_1 \) such that \( A \) causes \( A_1 \) and \( A_1 \) causes \( B \) (possibly a negative effect). Now, to calculate \( a \), we simply solve the equation \( P(B|A \leq a) - P(B) = 0 \), so:

\[
\Rightarrow P(B|A) - P(B) = 0 \\
\Rightarrow 10\frac{(89.925-30b+a(-8.9925+3b)+1.15629\times10^{-20}e^a)}{(-10+a)} - P(B) = 0
\]

Using an approximation method, we have found that for any value \( P(B) \), \( a = 10 \) and \( b = 3.88 \), but \( a \) cannot be 10 because \( P(A \leq 10) = 0 \) and \( P(B|A \leq 10) \) is indefinite. Thus \( A \) changes the probability of \( B \) for all the sample space in \( A \) for any value \( b \) in the sample space of \( B \).

Can we say that \( A \) causes \( B \)? No if the population is consider to be all the humans living in Puerto Rico for example. The reason is that the experiment was apply to only one human. In this case, it can be consider as a Token Cause which I’ll be explaining in details later. To consider this experiment as a causal structure, one of the things we have to make is to create multiple similar experiments for all the values of the manipulation variable, in other words, we have to make independent experiments with any possible population in existence. If the effect is the same to all experiments, then it’s more probable that the value given to the effect is a cause. If this is not the case, then there is a possibility that there exist an intermediate variable \( A’ \) between \( A \) and \( B \), whereas \( A \) cause \( A’ \) and \( A’ \) cause \( B \).

As you notice, I calculated the difference of the means between the distributions because it’s efficient for randomized experiment, in this example, it was sufficient to calculate the difference: \( |P(B|A) - P(A)| \). This value is a more precise way to find the causal structure in this example because it is only apply to me, but since we are willing to find a causal structure to all possible population in the area we are interested in, then the mean is a better way.

2.2 Bradford Hill’s Causal Criteria

I will explain some points that could helps us find the causal structure between the interested variables. It was established by Austin Bradford Hill, an English epidemiologist in 1965. These point are called the
Bradford Hill’s Criteria.

**Association Strength**

As we assume before, if there is a causal relationship between distributions, then they are correlated. Now, the association can be computed in several ways: find the correlation between variables, the regression coefficient in a regression system, or anything that is involve in changing the probability. As I said before, causal relationship between two variables implies correlation, but we cannot say the same the other way. Even if there is a strong correlation between variables, there is a chance that they are independent, for example, having a bad random data (rarely happens). Now, for randomize observational data sets, we have to be careful how we divide the population. Eventually, there can be false conclusion if we manipulate the causal variable badly. For example, consider the Simpson’s Paradox, it’s stated when for a given population, you divide it by specific groups (e.g. Male and female), you get different results when you apply the same experiment to the groups combine. Additional to this, there can be a correlation between variables when it is experimented with some specific group and no correlation at all when we work with the entire population.

**Consistency**

This is an important point to apply to experiments. This help us in analyzing the background knowledge between the interested variables and lower the probability in false conclusion in causal structures. As I mentioned above, the Simpson’s Paradox can help us to distinguish what kind of population we have to analyze so that we can construct the possibles (and true) causal relationship between the variables. If the effect appears in all the independent experiments that had been apply to all the possible divisions of the population, then there is a high chance that a causal connection between the possible causal variable and the effect exist. We have to consider all the possibilities for the manipulation variable; Type of object, time division (e.g. age division), condition of the object\(^2\), etc. Also, time has a great impact when we use consistency. Remember that our world (as we use as the sample space of observational data analysis) is in a constant change and all the time new objects (or statements) are created which can have an impact in the causal relationship of the interested variables. Suppose in Economics, before the internet was created, more specifically, online market was created, there was less impact in the global economics in general since there were less communication between the people around the world, so if you manipulate the economy in some certain country for example, then the impact in the economy in another country would be less probable.

\(^2\)By object, I refer to all possibles beings; humans, animals, material things, etc.
than when we manipulate the economy after the creation of the online market. So, considering applying the experiment in different intervals of time, we can guide the causal relationship to be a type-level (general) causal relationship if the conclusion is repeated in all possible time intervals, including the possible values of the manipulation variable that we are using. Moreover, if the Simpson’s Paradox is present then the groups we have divided may lead us to causal relationships, i.e., we have to get a deep analysis in the background knowledge for each group so that we can conclude the hypothesis we are testing. On the other hand, if the Simpson’s Paradox is absent for all division possibilities, then we can decide to follow two possibilities:

1) Apply the experiment to another set of population which is distinct (if exist) to the one we are analyzing and find if we got different results, if this is true, then we can state the Simpson’s Paradox an apply deep analysis as I mentioned above. If we got the same result, try to find another set of population, again distinct, and verify again the results given, and so on.

2) If 1) can’t be apply, i.e., we already analyze all the possible ways and still don’t get the Simpson’s Paradox statement, then we strive to analyze the entire population’s background knowledge in general and find the causal relationship. Notice, that this is the final (and only) results we can get, so there can be only one conclusion (if any), thus the causal structure we get is the only one for that randomized experiment in the interval of time. Remember that the world is changing constantly, so if there is a new object that appear in the sample space in the observational data, and has an impact in the experiment we are working on, then we have to re-do all the process again to see if there is a changing result.

In summary, the more specific we make the population, the better causal relationship we can find, this explains another point for the causal criteria

Specificity

The word says it all, the more you specify the population and restrict the values of variables studied in an randomized experiment, the possibilities of getting false causal structures are reduce. For example, if we like to studied how strong is the effectiveness of a certain medicine which cures a certain disease, then we apply the experiments to specific types of beings with the disease. Additional to that, we divide that specific population in subgroups of a specific characteristic which can be determine by the help of the background knowledge. After that, we verify if the Simpson Paradox can be mentioned and start some analysis of each group we manage to divide. Finally, check the results and come up with a possible causal structure.
Temporality

This is more of a logical statement: the effect has to occur after the cause. The cause variable has to be prior to the response. However, in randomized experiments, we are looking to increase (or decrease) the probability of the effect depending on our interest. Now, the background knowledge is the key to know which one can be the cause and the effect, since finding the "history" of each variable, we can denote, if possible, the time of discovery or creation of it. We can find logical statement or eliminate the possibilities of causal structure if we analyze well the available information so we can know which has to be cause and the effect (if there is a connection between them).

Biological Gradient

This is more for clinical trials experiments, this point states that the exposure of a variable leads to change the effect. The definition that I mentioned above in the beginning of this section can describe this point, for example, the more you increase the value of the cause variable, the greater the probability of the effect, if we are trying to nullify the effect, then the increase of the value of X leads to lower the probability of the effect.

Plausibility

This refers to an important key in causality, the background knowledge. This will help us minimize the error in jumping to false conclusions. Before starting any experiment of any subject, the first thing we have to do is to find all the information available about each interested variable and analyze it. This will help us to eliminate the misleading conclusions, as it is called in causality, the noisy data. Even though, we can’t eliminate them completely but analyze it independently and compare it with the same experiment without the noisy data, if there is a change then the noisy data can help us find the causal structure, if not then we can simply forget about them since they do not play an important role in the experiment.

Analogy

The effect of similar causes can be used to describe a causal relationship between the interested cause and effect. Sometimes other causes which are similar, may cause an effect which is in a deeper level, i.e., has an effect which can be an intermediate variable in the causal relationship between the interested variables.
Eventually, we analyze that intermediate variables using again background knowledge and try to associate it with the actual interested causal relationship we are working on. If there is no association at all, we discard that effect and continue on the others.

The other points are redundancy of the background knowledge definition. However, using the conditions will not imply that the causal structure we always get is the real one. In randomized experiment we measure the interested values most of the time, since we are 'illiterate' in a sense of what we are trying to conclude. Most of the time, we leave out variables that may have an impact in the causal structure. That’s why there exist many methods in statistics so we can be more sure that the conclusion we get are feasible. Also, the causal relationship often satisfy the Markov Condition\(^3\) and faithfulness condition\(^4\). The program that I will use (TETRAD) use these conditions to create the possible causal model. Finally, we have some necessary conditions that will help us find a minimal-error causal structure or the nearest real causal structure. On the next section, I will explain the types of causal structures and how and when to apply them in randomized experiments.

3 Types of Causalities

3.1 Introduction

There exist many types of causalities which can be use depending of what experiment we are working for (social, biological, etc.). Some types may not seem useful in some of the cases, like cyclic causalities in clinical trials, but anyway, we have to consider it just in case there is some missing information. Token causalities are often use in a single randomized experiment of a specific population. Type-level causality are use to conclude the conjunction of each Token causality found of the same randomized experiment given for all possible population. We can say that the type-level causality is the general causal structure of the interested variables. Remember, one of the points of the criteria is that if the effect is repeated often in multiple independent randomized experiments, then there is a chance that there is a causal relationship. Also, I’m going to explain the strength of the causal relationship between variables and how this is going to help us in randomized experiment.

---

\(^3\)Markov conditions states that for each \(X\) in the causal structure, denoting \(PA_X\) and \(ND_X\) being the parents of \(X\) and non-descendants of \(X\) respectively, then \(X\) is conditional independent of \(ND_X\) given \(PA_X\), i.e.: \(I_P(X, ND_X|PA_X)\)

\(^4\)Faithfulness conditions states that all conditional independence has to come from the Markov Condition and only Markov Condition.
3.2 Token Causality

As I mentioned above, the example given can be consider as a token causal structure, which as a simple description, it refers to the causal structure to a particular occasion, in this case, for me (for more details, please read Kleinberg 2013). In statistics, a token causality can be consider if it was the first experiment made of a particular set of variables. If we repeat the same experiment, but for a particular population distinct of the previous one and get another causal relationship, with all the background knowledge analyze, then it’ll be also consider as another token causality. Finally, we continue with the same process until we satisfy one of the points in Bradford Criteria. Now, to consider a good token causal relationship in a experiment, we have to observe the following, before stating the consideration, I will define the following:

Let $U_O$ be the set of all object existence in the universe which we are experimenting on (e.g. human, animal, etc.) and let $C_O$ be the set of all characteristics of all existent object could have. By characteristics, we include their physical, emotional and cultural description. Now a subset $P \subset C_O$ contains the characteristics of a particular object could have. So for a randomized experiment, the subject we are experimenting is an element $o \in U_O$ and the variables in the experiment are consider the characteristics of that object, i.e., the elements $c \in P$. Each element of $P$ is a vector of $n$ values, where $n$ is determine by the number of object studied in the randomized experiment and the values are given. For example, suppose we get the following data from 15 human patients:
<table>
<thead>
<tr>
<th>BP(systolic)</th>
<th>BP(diastolic)</th>
<th>Glucose</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>80</td>
<td>75</td>
<td>168</td>
<td>68</td>
</tr>
<tr>
<td>135</td>
<td>85</td>
<td>110</td>
<td>170</td>
<td>78</td>
</tr>
<tr>
<td>145</td>
<td>90</td>
<td>125</td>
<td>163</td>
<td>82</td>
</tr>
<tr>
<td>150</td>
<td>95</td>
<td>141</td>
<td>159</td>
<td>92</td>
</tr>
<tr>
<td>125</td>
<td>80</td>
<td>117</td>
<td>171</td>
<td>81</td>
</tr>
<tr>
<td>120</td>
<td>85</td>
<td>91</td>
<td>169</td>
<td>75</td>
</tr>
<tr>
<td>90</td>
<td>60</td>
<td>78</td>
<td>164</td>
<td>66</td>
</tr>
<tr>
<td>160</td>
<td>110</td>
<td>138</td>
<td>157</td>
<td>89</td>
</tr>
<tr>
<td>130</td>
<td>85</td>
<td>103</td>
<td>170</td>
<td>77</td>
</tr>
<tr>
<td>125</td>
<td>85</td>
<td>88</td>
<td>170</td>
<td>74</td>
</tr>
<tr>
<td>140</td>
<td>95</td>
<td>86</td>
<td>171</td>
<td>79</td>
</tr>
<tr>
<td>130</td>
<td>85</td>
<td>93</td>
<td>167</td>
<td>82</td>
</tr>
<tr>
<td>150</td>
<td>100</td>
<td>137</td>
<td>165</td>
<td>88</td>
</tr>
<tr>
<td>135</td>
<td>90</td>
<td>101</td>
<td>161</td>
<td>79</td>
</tr>
<tr>
<td>110</td>
<td>75</td>
<td>95</td>
<td>179</td>
<td>82</td>
</tr>
</tbody>
</table>

Where the variables are:

**BP(Diastolic)** = Diastolic Blood Pressure  
**BP(systolic)** = Systolic Blood Pressure  
Glucose = Glucose levels in mg/dL  
Height = Height in cm  
Weight = Weight in kg

So the subject $o = human$ is consider to be an element in $U_O$ and the variables mentioned above are the characteristics of that object. Every variable in $P$ has a value, and each interval of values of each variable depends of its distribution, so $P$ in this case contains 15 elements, which are the values given in that table. Thus, each variable has a value for each human. Now, I will introduce a set $O^*$ where the elements in that set will depend of what variables we are interested in finding a causal relationship, i.e., the specific population. In this example, if we are interested in finding a causal relationship in Glucose and Weight (remember that we have to find first all possible correlation between all the variables to have a good division of specific popu-
lation), then a possible element \( P^* \in O^* \) can be define as \( P^* = \{ p \in P | 75 < p(Glucose) < 110 \} \). As you can see, the elements of \( O^* \) are sets of objects with the characteristics specified by one, and \( \bigcup_{i=1}^{m} P^* = P \), where \( m \) is the number of elements in \( O^* \). However, the separation between specific population are done using the background knowledge so we can get the desire results. Using this method, we can state the Simpson Paradox in observed experiments, since we divide the population into a more specific one with certain characteristics. Now, suppose that you got a causal relationship between the interested variables \( X \) and \( Y \), where \( X \) is the cause of \( Y \) with the specific population \( P^* \), then a good token causality is consider if for a subject \( o \in U_O \):

\[
a \leq P(Y) - P(Y|X) \leq b \text{ for } |P^*| \rightarrow \infty
\]

In other words, the change in probability is bounded between two numbers, this is because of various reasons, one of them is the errors in values since we are assuming that the data is normalize. If \( a - b = 0 \) then it is consider a strong causal relationship, moreover, a “perfect” causal relationship. On the other hand, if \( |a - b| > 0 \), then it’ll depends of the background knowledge of the response variable and the statistician hypothesis. For example, if there’s a change of a characteristic of the object we are studying, which is consider in the hypothesis we are testing, then we can consider it as a good causal structure. What I mentioned above is the strength of the relationship, the smaller the difference \( |a - b| \) is, the greater the causal relationship between them.

### 3.3 Type-level causality

Now, after various experiments made, we have to find if there’s a relationship between them, i.e., compare all token causality found in all independent experiments and find if there is a repetition of connections between the interested variables, if so, then we can consider the causal relationship as Type-level causality. Type-level causality is a causal relationship that occurs in general. In general, for all the characteristics in \( C_O \) for a particular object in \( U_O \), the causal relationship between the interested characteristics (variables) is present for all possible experiments made by the statistician, then it can be considered as a Type-level causality. For example, suppose that for an object \( o \in U_O \) we get for each experiment the following Token Causal graph:

---

5. \( c_o[i] \) is the \( i \)th element of the vector, for \( c_o[Glucose] \) is the value of glucose.

6. \( |P^*| \) = cardinality of \( P^* \)
then it can be consider as a Type-level causal graph. In other words, we manage to experiment the subject (object) in all possible ways, i.e., we studied the object with all the characteristics in $C_O$ available for it. If there is a token causal graph which the interested variables are differently connected to the others token causal graphs, assuming that there are no errors, then:

1. If there is a characteristic in that particular graph which is not in the others, then we have to consider it as an interested variable.

2. if with 1), the problem still persist, then there exist a hidden characteristic which we didn’t consider it in the experiment.

4 Causal Strength

4.1 Spurious vs Genuine

Before explaining the comparison between spurious and genuine cause, I will explain what is a “Prima Facie cause”. A Prima facie cause according to Kleinberg 2013 has to satisfy the following conditions:
1. For two states (or distributions) X and Y, to X be a possible cause of Y, X has to be a temporary prior to Y, i.e., for a certain time X cause Y.

2. X must change the probability of Y.

Now, X is a $\epsilon$-spurious cause of Y if X is a prima facie of Y and for W which is prior to X we have:

1. $P(X) > 0$
2. $P(X \cap W) > 0$
3. $|P(Y|X) - P(Y)| > 0$
4. $|P(Y|X \cap W) - P(Y|W)| < \epsilon$

I put an absolute value in “3.” because X can be a negative cause of Y. Now a genuine cause is the cause which "dominates" the other possible causes of Y, i.e., is a non-spurious cause of Y. In randomized experiments, the characteristic which change the probability of the interested response at most, is the one which can be considered as a genuine cause. Now, a way to calculated a possible genuine cause, we can use the formula that calculate the significance of a cause X to a response Y, which is defined as:

$$\left| \sum_{c \in C/X} P(Y|X,c) - P(Y|X^c,c) \right|_{|C/X|} < \epsilon$$

(2)

This is a general formula for all types of causal relationship. Kleinberg uses a more specific one for Token causal relationship, she compares it using as prior knowledge the type-level causal relationships of similar experiments done before. Also with interval of times, considering the change of characteristic depending of what time and place is done the experiment, adding to (1) the probability of the cause given the observed prior experiments and the weighted function $f(x, y, r, s)$ where $r, s$ are time variables such as $0 < r < s < \infty$. However, my interest is in using characteristics of the subject study to conclude the causal relationship between the interested variable, despite of time. My approach is similar to using time, but I focus more on considering the types of population we are working, i.e., consider all the characteristic in existent of the particular object we are working. So, the insignificant causes are considered as interested too because, there can be a connection between the spurious and the non-spurious cause, with more chance that
the non-spurious cause is an intermediate variable between the spurious one and the interested response. For example, suppose that we have two chemical liquids G1 and G2 with each 5 mg of substances in which we want to know the effect in human cell regeneration in muscles. Suppose also that we study 300 human subjects with identical interested characteristics and the interested response characteristic Y which is: cell regeneration rate in percentage and suppose also by background knowledge that the mean of Y in human being with the interested characteristic is 15 with normal distribution. This population is divided in 3 independent experiment in which we tested 100 human subjects who exercise regularly without G1 nor G2, 100 human subjects who exercise regularly given G1 and 100 human subjects who exercise regularly with G2 and we get the following PDF of each sub-population given:

\[
f(y|15, 3) = \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15)^2}, \quad -\infty < x < \infty
\]

\[
f(y|15, 3, G1) = \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15-2G1)^2}, \quad -\infty < x < \infty
\]

\[
f(y|15, 3, G2) = \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15-4G2)^2}, \quad -\infty < x < \infty
\]

Now, for G1 and G2 be a cause of Y in this case, the following has to satisfy for \(Gi, i = 1, 2\). Letting \(y^*\) the value of the biggest solution \(y\) of the equation \(f(y|Gi) - f(y) = 0\) (considering that G1 and G2 are positive causes):

1) \(P(Y > y^*) \approx 0\)

2) \(P(Y > y^*|Gi) \approx 1\)

It can be shown that 1) leads to 2). The reason of these conditions is that we have to maximized the quantity of the studied population such that their interested response characteristic change, since if \(P(Y > y^*) > 0\) then:

\[P(|Y - y^*| < m) = P(|Y - y^*| < m|Gi)\]

for \(m > 0\) and \(\sigma_Y = \sigma_Y|Gi\)

More general (\(\sigma_Y \neq \sigma_Y|Gi\)), we have that

\[P(y^* - m_1 < Y < y^* + m_2) = P(y^* - m_1 < Y < y^* + m_2|Gi)\]
For some $m_1 > 0$ and $m_2 > 0$

In other words, the causal variable doesn’t change the probability of the response implying that the second condition of a Prima Facie cause doesn’t satisfy. Now using the conditions above we calculate how much of mg do we need of G1 and G2 so that the response characteristic change for all the specific population. For G1 we have to solve the following:

\[
f(y|15, 3, G_1) - f(y|15, 3) = 0
\]

\[
\Rightarrow \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15-2G_1)^2} - \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15)^2} = 0
\]

\[
\Rightarrow \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15-2G_1)^2} = \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15)^2}
\]

\[
\Rightarrow e^{\frac{1}{18}(x-15-2G_1)^2} = e^{\frac{1}{18}(x-15)^2}
\]

\[
\Rightarrow \log(e^{\frac{1}{18}(x-15-2G_1)^2}) = \log(e^{\frac{1}{18}(x-15)^2})
\]

\[
\Rightarrow \frac{1}{18}(x-15-2G_1)^2 = \frac{1}{18}(x-15)^2
\]

\[
\Rightarrow (x-15-2G_1)^2 = (x-15)^2
\]

\[
\Rightarrow x = 15 + G_1
\]

Now we want to make $P(Y > 15 + G_1|G_1) \approx 1$ and $P(Y > 15 + G_1) \approx 0$. I will use a tolerance equals to $10^{-7}$. So solving:

\[
\int_{15+G_1}^{\infty} \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18}(x-15-2G_1)^2} dx = 0.9999999
\]

We have that $G_1 = 15.598mcg$

Now for G2, doing the same process we did with G1, we got that $G_2 = 7.7901mcg$. So in conclusion,
we got that G2 is stronger that G1, using (1) for both causes, and suppose that letting $G_1 = G_2 = 7.79901$ we want to calculate how strong is . we have that:

$$\Rightarrow \left| \sum_{c \in C/X} P(Y|G_1, G_2) - P(Y|G_1^c, G_2) \right| \left| \frac{|C/X|}{|C/X|} \right|$$

$$\Rightarrow \left| P(Y|G_1) - P(Y) \right|$$

$$\Rightarrow \left| \int_{40}^{50} \left( \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15-2 \times 7.79901)^2} - \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15)^2} \right) \right|$$

$$\Rightarrow 0.000862224$$

Now for G2 we have that:

$$\Rightarrow \left| \sum_{c \in C/X} P(Y|G_2, G_1) - P(Y|G_2^c, G_1) \right| \left| \frac{|C/X|}{|C/X|} \right|$$

$$\Rightarrow \left| P(Y|G_2) - P(Y) \right|$$

$$\Rightarrow \left| \int_{40}^{50} \left( \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15-4 \times 7.79901)^2} - \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15)^2} \right) \right|$$

$$\Rightarrow 0.878154$$

Moreover, for any rate of cell re-generation, i.e., $0 < y < a$ where $a > 0$ we have that for G1:

$$\Rightarrow \left| \int_{0}^{a} \left( \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15-2 \times 7.79901)^2} - \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15)^2} \right) \right|$$

$$\Rightarrow 2.86652 \times 10^{-7} + 0.5Erf[3.53553 - 0.235702a] - 0.5Erf[7.21202 - 0.235702a]$$

and for G2:

$$\Rightarrow \left| \int_{0}^{a} \left( \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15-2 \times 7.79901)^2} - \frac{1}{3\sqrt{\pi}} e^{\frac{1}{18} (x-15)^2} \right) \right|$$
We have the following plot, where $a$ is the x-axis:

As you can see, if $G_2$ is given, it tends to significantly change the response variable more than $G_1$ as the interval $0 < y < a$ grows. Concluding that $G_2$ is a stronger cause than $G_1$. Also, for any interval $b < y < a$ we have the following plots:
As you can see, the yellow shade area where $P(Y|G1) > P(Y|G2)$ is the area where $P(Y|G1) \approx 1$ and $P(Y|G2) = P(Y) \approx 0$. However, most of the time, $P(Y|G2) > P(Y|G1)$ meaning that $G2$ is a stronger cause than $G1$. A causal graph can be as the following:
Since G1 and G2 is given with exact amount to the population, and G1 is a spurious cause, then a) is possible, i.e. it is possible that G1 cause G2 and G2 causes Y with a possibility that G2 is a chemical structure which is in G1 \((G1 \cap G2 = G2)\). Another possibility is that there is a characteristic in common that G1 and G2 have, which change the probability of Y as it is in b). Now there exist more causal graph that we can create using the experiment made, if we include a population such that are given G1 and G2, but this can be consider if there exist a reaction between a substance of G1 with another substance of G2, we can know this we background knowledge.

5 Method to reduce error in finding causal relationship for Experimental studies using data sets

Here I will present a method so that we can get the most accurate causal graph given a data set. Sometimes statisticians predict false conclusion with data sets, since these are just random samples taking of a random population. In causal prediction, we have to pick the population carefully such that it has the sufficient and necessary characteristics of the subject we are studying, so that the interested variables get the most accurate causal relationship. I cannot call this method a definition since we haven’t got any method to study the root of the material, for example, it is difficult (nearly impossible) to study the atoms structure in any subject, that would take too much time and even error data depending how we are measuring, but we would get the optimal causal graph. With the help of Bradford Hill’s Criteria, the following steps shows how to get an optimal causal graph:

1) **Background Knowledge**. First, analyze well the interested variables which we want to find the causal
relationship. Search for definitions about the variables or definition which are a connection to them and past studies about them if available. This will help us how to divide the population and what characteristics should be considered in the experiment.

2) **Simpson Paradox.** Always check if for the given definition of the interested variable, dividing the population into sub-populations with specific characteristic values we get different results. For example, a characteristic of a human can be the gender, which differ in a significant number of other characteristics like body parts, hormones levels, etc. Also the age is very crucial here, since the human body tends to change as we grow (again hormones levels, body parts, etc), so their characteristic values changes too.

3) **Association between the interested variables.** After you divided the population, find the correlation between the variables. If it has a weak correlation, try to divide the population in more specific groups using step 1) of this method. It suppose to get a stronger correlation every time we specify the population. In a mathematical sense, the manipulation variable has more values to select and create the causal graph, this happens because we get closer to the root of the variables, the origin of them. However, keep in mind that if the variables are correlated, it doesn’t mean that they are causally related.

4) **Use TETRAD program for searching pattern and optimal regression equations to create a good causal model.** This program will use the PC algorithm to search a model which satisfies the Markov Causal Assumption such that it fits the statistical data given, with the chi-square fitness, it’ll facilitate us the computing process of finding the correlation, regression coefficient, etc.

5) **Conclusion and future analytic process for the same experiment.** After finding the optimal causal relationship between the variables, analyze and conclude the reason of the given relationship. However, we have to make multiple experiments for the same subject. In addition for all characteristic available for that subject which has some connection with the interested variables so that we get the true causal graph for the interested experiment.

### 6 Experiments

#### 6.1 Causality in patient’s diagnostics on Glucose level

Consider the following example:
15 patients were diagnosed to verify the glucose level, the blood pressure level when the heart has a contraction (systolic), the blood pressure level when the heart is relax (diastolic), the following data was obtained:

<table>
<thead>
<tr>
<th>BP(systolic)</th>
<th>BP(diastolic)</th>
<th>Glucose</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>80</td>
<td>75</td>
<td>168</td>
<td>68</td>
</tr>
<tr>
<td>135</td>
<td>85</td>
<td>110</td>
<td>170</td>
<td>78</td>
</tr>
<tr>
<td>145</td>
<td>90</td>
<td>125</td>
<td>163</td>
<td>82</td>
</tr>
<tr>
<td>150</td>
<td>95</td>
<td>141</td>
<td>159</td>
<td>92</td>
</tr>
<tr>
<td>125</td>
<td>80</td>
<td>117</td>
<td>171</td>
<td>81</td>
</tr>
<tr>
<td>120</td>
<td>85</td>
<td>91</td>
<td>169</td>
<td>75</td>
</tr>
<tr>
<td>90</td>
<td>60</td>
<td>78</td>
<td>164</td>
<td>66</td>
</tr>
<tr>
<td>160</td>
<td>110</td>
<td>138</td>
<td>157</td>
<td>89</td>
</tr>
<tr>
<td>130</td>
<td>85</td>
<td>103</td>
<td>170</td>
<td>77</td>
</tr>
<tr>
<td>125</td>
<td>85</td>
<td>88</td>
<td>170</td>
<td>74</td>
</tr>
<tr>
<td>140</td>
<td>95</td>
<td>86</td>
<td>171</td>
<td>79</td>
</tr>
<tr>
<td>130</td>
<td>85</td>
<td>93</td>
<td>167</td>
<td>82</td>
</tr>
<tr>
<td>150</td>
<td>100</td>
<td>137</td>
<td>165</td>
<td>88</td>
</tr>
<tr>
<td>135</td>
<td>90</td>
<td>101</td>
<td>161</td>
<td>79</td>
</tr>
<tr>
<td>110</td>
<td>75</td>
<td>95</td>
<td>179</td>
<td>82</td>
</tr>
</tbody>
</table>

Now analyzing the data, we have the following plots shown in Figure 1 between $X$ vs $Y$, $X \neq Y$, for all $X,Y$ in the data.

First of all, I will consider weakly correlated of both variable if $R^2 < 0.50$. As you can see, Glucose and height (Fig. 7) has a weak correlation since $R^2 = 0.279$, so we can say that Glucose and Height are causally disconnected or that there are insignificantly connected. Now for BP(diastolic) vs Glucose (Fig. 1), it has the same problem as Glucose vs Height, so we can leave out that correlation as well. As for the others, there exist a significantly correlation between them.

**Hypothesis**

First let start with Glucose and Weight. As you can see, they are well-correlated. However, from background knowledge, the “weight” variable is exogenous in this set of data, therefore weight can’t be an effect, so it leave us that it can be a possible cause for this experiment. Secondly, I can’t be sure if weight causes
Figure 1: Correlation between the variables
glucose since we can develop weight in different aspects, like for example; Muscle weight (bulking your body), eating excess of fat, or eating carbs. Since they are correlated, I assume that there is a common cause which isn’t in the data. Now, about BP (diastolic) and weight, I can say the same as glucose and weight. The blood pressure can be cause by high energy in your body (ex. adrenaline), which is different compare to weight, so, I will think there is a common cause between them, maybe a weak one since $R^2 = 0.563$. Both blood pressure variables has a strong correlation, which means that there is a possibility that one cause another, or viceversa, but it is illogical that one cause the another since they are independent from each other, so I will also assume that there is a common cause too. Finally, for the other correlations, I will assume they weakly connected.

Using the TETRAD program, I will run the FCI Algorithm to find the patterns between the variables, and I got the following:

As you can see, my hypothesis is almost true. Notice that neither diastolic and systolic are causally connected with weight. Now, both pairs; Glucose and weight and BP(diastolic) and BP(systolic), are independent. Also, Glucose and Weight are causally connected, but, we can’t say if one cause another since the circles that are in the figure above, indicates that it can have an arrow or not, so there are three possibilities:
Figure 3: Possible causal models: the numbers in green in this case are the variable’s mean and the black one are the regression’s coefficient.

1. Glucose cause Weight

2. Weight cause Glucose

3. Neither one cause the other, which means, that there is a hidden common cause between them.

I will estimate all possible model to verify which is the most suitable to cause one another. Notice that the FCI algorithm already estimates all possible models to verify which is the best, but for sake of specification, I will manually estimate each model to see the statistics details. The following models in Figure 3 shows the possible causal graphs:

When I estimate the two models, I got the following information shown in Figure 4:

As you can see, the P value for both models are 0, which means that the data doesn’t fit the model in the cases assume. For BP (diastolic) and BP (systolic), the models have the same model statistics’ details. Therefore, the only possibility that the model may fit the data is that we introduce a hard intervention variable, which is strongly correlated to both variables (or four). So, I can’t construct a Causal Bayes model from the given data since we can’t find a model that fit the data statistically.
Figure 4: *Statistical data fitness*

Degrees of Freedom = 8  
Chi Square = 355.0775  
P Value = 0.0000E0  
BIC Score = 333.4131  
CFI = 0.9581  
RMSEA = 1.7604

The above chi square test assumes that the maximum likelihood function over the measured variables has been minimized. Under that assumption, the null hypothesis for the test is that the population covariance matrix over all of the measured variables is equal to the estimated covariance matrix over all of the measured variables written as a function of the free model parameters—that is, the unfixed parameters for each directed edge (the linear coefficient for that edge), each exogenous variable (the variance for the error term for that variable), and each bidirected edge (the covariance for the exogenous variables it connects). The model is explained in Bollen, Structural Equations with Latent Variable, 110. Degrees of freedom are calculated as $m (m + 1) / 2 - d$, where $d$ is the number of linear coefficients, variance terms, and error covariance terms that are not fixed in the model. For latent models, the degrees of freedom are termed 'estimated' since extra constraints (e.g. pentad constraints) are not taken into account.
6.2 Causality in patient’s coronary arteries disease

I want to study whether or not the level of cholesterol cause severe coronary disease through duration of coronary disease symptoms. This dataset is from the Duke University Cardiovascular Disease Databank and consists of 3504 patients and 6 variables (I will leave out sex, and age as exogenous variables). The patients were referred to Duke University Medical Center for chest pain. I will divide this data set from patient with ages:

a) 17 - 35
b) 36 - 50
c) 51 - 65
d) 66 and older

Let start out with a), using the sub-data set, I got the following correlation between the variable Cholesterol vs Duration of symptoms of CAD shown in Figure 5.

As you can see, there is no correlation whatsoever between the duration of CAD symptoms and cholesterol, so we can say that they are independent. Now let assume that the variables of this data set has a normal distribution, and let consider as they are correlated linearly (so the causal graph can be more consistent). The Significant Coronary Disease by Cardiac Catherization (SCDCC) variable has two values: SCDCC = 0 as “no” and SCDCC = 1 as yes, same as the Three Vessel or Left Main Disease by Cardiac Catherization (TVLMDCC). Now, I found the folowing:
\[
\beta_{01} = E[\text{cholesterol}|SCDCC = 0] = 207.29
\]
\[
\beta_{02} = E[\text{cholesterol}|SCDCC = 1] = 248.24
\]

As you can see, people with SCDCC have higher level of cholesterol on average than the people who don’t have SCDCC. So we can say that cholesterol and SCDCC have a correlation between them. Now, let use TETRAD to see the connection between the variable, we know that TVLMDCC is the predictor of SCDCC, and that Cholesterol and Duration of CAD symptoms are un-correlated, thus they are independent to each other. Using that background knowledge, and the FCI algorithm, I got the following model shown in Figure 6.

![Possible causal graph](image)

As explained before, we have multiple possibilities for this graph. Analyzing the data, I got two models that fits the data with their higher P-value respectively shown in Figure 7.

Seeing this, the duration of CAD symptoms is independent of cholesterol and any other available. However, we cannot be sure if SCDCC cause cholesterol or they have a common cause. As I know, having a significant CAD doesn’t cause cholesterol since you can get high level of cholesterol by consuming too much fat. So the best model that fits the data between patients from ages 17-35, is the second one shown above.
Figure 7: Optimal causal graph with their p-value (high p-value indicates that the model fits good, so the higher the p-value, the better is the causal graph)
Now for the patients from 36-50 in Figure 8, I got a very weak statistical model, i.e., the best model that best fit the data, which is shown in figure 9 has a P-value = 0.0232 < 0.05 so the hypothesis rejects that model, even it is almost equivalent to the sub-data set of 17 - 35 years old patients, we cannot predict if cholesterol causes SCDCC.

Now for the patients 51-65, using the FCI algorithm, in figure 10 shown the result. This is interesting, the first two sub-data set, we got that the duration of CAD symptoms was independent to the other variables. Now, this model tells us that cholesterol has a correlation with the duration of CAD symptoms, so there are two possibilities:

1. That the level of cholesterol cause that CAD symptoms tends to last longer

2. That there is a common cause between them, maybe there something that is causing to have influence cholesterol and the duration of CAD symptoms (for example, eating too much junk food)

Moreover, there is a chance that having a severe CAD disease causes the duration of CAD symptoms last longer by common sense or there is a common cause between them.

Analyzing the distinct types of possible models, the best model that fit the data is shown in figure 11 with its p-value respectively.

As you can see the P-value is high, so this model fits well with the data given.

Now for the patients with ages from 65 and older, using the FCI algorithm, figure 12 shows the following
Figure 9: Optimal causal graph for 17-35 years old patients with its p-value respectively.

Figure 10: Possible causal graph for 51-65 years old patients.
Estimating the different sub-data sets, we got that for patients between 17 and 50, having higher cholesterol doesn’t influence the duration of CAD symptoms, but has a common cause with SCDCC. On the other hand, we have that for patients from 51 and older, the cholesterol and SCDCC doesn’t have a correlation whatsoever, but cholesterol has a common cause with the duration of CAD symptoms. I assume that there is some intervention variable, which cause all variables for the two cases (17-50 and 51-older) such that makes both models into one. Maybe it is something about food or exercise, since when you are younger, you are more active, you exercise more and you process the food better. That’s why you can “nullify” the duration of the symptoms since you are “curing” yourself with exercise. However, when you are older, you are less active (probably most of the people) so if you consume food that make you rise your levels of cholesterol,
you duration of CAD symptoms tends to last longer since you are less active, implying that you can’t “cure” yourself without using any other medication (where this is some other case).

6.3 Causality in Prostate cancer 1985 study

Byar and Greene prostate cancer data, from Andrews DF and Herzberg AM (1985): Data. New York: Springer-Verlag, and lib.stat.cmu.edu) with 502 patients observations made. The following variables are included in the experiment:

\[ \text{stage} = \text{Cancer Stage} \]
\[ \text{rx} = \text{Medication (Manipulation Variable) with the following values:} \]

1. Placebo
2. 0.2 mg of estrogen
3. 1 mg estrogen
4. 5 mg of estrogen

\[ \text{status} = \text{status of patient with values:} \]

1. 1 = alive
2. 2 = dead - cerebrovascular
3. 3 = dead - heart or vascular
4. 4 = dead - other calcification
5. 5 = dead - other specific non-calcification
6. 6 = dead - prostatic calcification
7. 7 = dead - pulmonary embolus
8. 8 = dead - respiratory disease
9. 9 = dead - unknown cause
10. 10 = dead - unspecified non-ca
age = Age in Years

Weight Index = weight(kg) - height(cm) + 200

pf = patient activity with values:

1. 1 = in bed < 50 percentage of the time in day
2. 2 = normal
3. 3 = in bed > 50 percentage of the time in day
4. 4 = confined to bed

History of Cardiovascular disease = values:

1. yes = 1
2. no = 0

Systolic Blood Pressure/10 = Ex. 1f 120 mmHG then 12 will be on the value.

Systolic Blood Pressure/10 = Ex. 1f 120 mmHG then 12 will be on the value.

ekg = Electrodiagram results with values:

1. 0 = benign
2. 1 = heart block or conduction def
3. 2 = heart strain
4. 3 = normal
5. 4 = old MI
6. 5 = recent MI
7. 6 = rhythmic disturb and electrolyte ch

Serum Hemoglobin (g/100ml)

Size of Primary Tumor (cm²)

Combined Index of Stage and Hist. Grade

Serum Prostatic Acid Phosphatase

Bone Metastases = values:

1. yes = 1
My interest in this experiment is to check if there is a change of patient’s characteristics given an amount of estrogen hormone. As you can see, the levels of estrogen dosage to the patient are well distributed, it can be seen that is exponentially increased (i.e., $5^x$ mg of estrogen for $x = -1, 0, 1$). However, this experiment is not created to find a causal graph, so I can’t guarantee a good causal graph, instead I would find the optimal one for this experiment but it doesn’t mean that it is the true causal graph since there aren’t sufficient characteristic so that we can divide the population the way I want. Now, I wouldn’t divide the population more since the age interval is between $48 < a < 84$ which I don’t consider that there is a significant change in patient characteristic. Furthermore the weight index isn’t well explained so I can’t find information which I can analyze and divide the population such that it has a correlation with the prostate cancer study. So the only values for the manipulation variable are the ones given above. Using this background knowledge, and TETRAD program to find the causal graph optimization with the PC algorithm for Markov Causal Assumptions, the figure 13 and 14 shows the respectively graph for each specific population.

Using the background knowledge, the following graphs in figures 15 and 16 shows their optimal causal relationship between the variables for all the values in the manipulation variable with the their estimation.

Here you can see that there isn’t a single causal connection which you can see that it increase (or decreased) its coefficient variable as a pattern. If you compare each optimal graph there are so many changes that you can’t find at least a logical pattern between those changes, for example, the age characteristic is an independent variable for the population given 0.2 mg of estrogen and 5 mg of estrogen. However, for the population given placebo (0 mg of estrogen) and 1 mg of estrogen the age characteristic change the probability of 2 distinct response variables, which those response variables don’t have a correlation between them. Thus, I can conclude that there are missing characteristics in the experiment which are important, i.e., it’s needed to divide the population such that we can find a good causal graph. On the other hand, the Systolic and Diastolic have the same direction for all the manipulation variables (as expected) still we can’t say that Diastolic causes Systolic since the regression coefficient between those variables changes without a pattern as you can see above in the graphs. Again, we need more characteristics (variables) included in the experiment.

Something interesting we have here. The size of the primary tumor has a correlation with Bone Metastases for the patients given placebo, stating that Bone metastases can determine how the size of the primary tumor is, but as the population is given a quantity of estrogen, those characteristics are independent. By definition estrogen hormone have multiple effects in growth and development, and one of the target is the bone marrow, which is the preferred metastatic site for the prostate cancer tumor. On the other hand the direction changes

2. no = 0
Figure 13: Possible causal graphs: a) patients given placebo; b) patients given 0.2 mg of estrogen
Figure 14: Possible causal graphs: c) patients given 1 mg of estrogen; d) patients given 5 mg of estrogen;
Figure 15: Optimal causal graphs: a) patients given placebo; b) patients given 0.2 mg of estrogen
Figure 16: Optimal causal graphs: c) patients given 1 mg of estrogen; d) patients given 5 mg of estrogen;
from Bone Metastases and Serum Prostatic Acid Phosphatase as its increments the value of mg of estrogen given to the population. There are many possibilities: The hormone levels of each patient are slightly different for each one, since the hormones changes depending how’s your diet or what other medication he’s taking, etc. So in conclusion, too many factors can be consider here, that’s why, like I said before, that we need to measure more characteristics of the patients to consider it for the experiment, specifically, ones that are connected with the estrogen hormone.

6.4 Causality in Prostate Cancer between patients in the USA and PR

Same as the experiment above, I will divide the population between cases (case or control) and for each population, I will divide through ages to verify if there is some changes through causal graphs between the specific populations. First I want to state that this experiment is not complete since we have significantly missing values, so the optimal causal graph will not be a good one. The following characteristics (variables) are available for this experiment using as subject the human:

**Age diabetes** = values

1. yes = 1
2. no = 0

**heartdisease** = values

1. yes = 1
2. no = 0

**stroke** = values

1. yes = 1
2. no = 0

**Blood Pressure** = values

1. yes = 1
2. no = 0

cigarrate = Smoke cigarettes, values

1. current smoker = 1
2. Previous smoker = 2
3. never smoked = 3

weightinlbs = weight in pounds
heightin = height in inches
precontrolserum = PSA level
metastasis = values

1. yes = 1
2. no = 0

prostatectomyweightgrams = Prostatectomy weight in grams gleason i, i = 1,2,3 = gleason i grade
totaltumorvolume = Total tumor volume (cc) stagetumor = stage of the tumor (1,2,3,4)
lymphnode = Lymph Node stage
wa = African Ancestry
ea = European ancestry
na = Native ancestry
SNP = Number of alleles present from genes related to prostate cancer risk

Given the division and using Background knowledge for the given characteristics, we get the following graphs using the PC algorithm:

For the case population, dividing the population thorough the age intervals: 40-60, 60-70, 70-90 we have the following graphs shown in figure 17 and 18 respectively.

For the control population, dividing the population thorough the age intervals: 40-60, 60-70, 70-90, the figures 19 and 20 show the results.
Figure 17: Possible causal graphs: a) 40-60 years old case patients; b) 60-70 years old case patients

Figure 18: Possible causal graphs: c) 70-90 years old case patients
Figure 19: Possible causal graphs: a) 40-60 years old control patients; b) 60-70 years old case patients

Figure 20: Possible causal graphs: c) 70-90 years old control patients
Figure 21: Optimal causal graphs: a) 40-60 years old case patients; b) 60-70 years old case patients

Figure 22: Optimal causal graphs: c) 70-90 years old case patients
Figure 23: Optimal causal graphs: a) 40-60 years old control patients; b) 60-70 years old case patients

Figure 24: Optimal causal graphs: c) 70-90 years old control patients
Analyzing the following graphs using the information of each variable, we have the following optimal causal graph with the given regression coefficients:

Observing and comparing the graph, it’s difficult to conclude an optimal causal graph for the following reasons:

1. **Insufficient data.** As I said before there are many missing important values (more than 50%) which affect too much that analysis of causal relationship between the interested variables.

2. **Values of the variable.** Many variables are discrete and other continuous. Moreover, the variable which are discrete, are very generalized which it make the experiment for difficult to divide, as for example, Blood Pressure variable, the only values they had are yes or no, instead of using those values, use the blood pressure value of the patient, first because it is more specific and secondly, it’s continuous (better to normalize).

Furthermore, dividing the population which include only the patients who has a prostatectomy check (i.e. with the minimum missing values available) we have the following optimal causal graphs with their estimator respectively shown in Figure 21 and 21.

It’s interesting that the gene alleles has a correlation with the diabetes in the case patients, but since the p-value is very low, and the Chi-square error is high, then we cannot conclude anything since this optimal causal graph it’s not good.

### 6.5 Causality of USA and PR Prostate Cancer data using Ang Li’s Statistical analysis to find the significant variables between specific population

Ang Li uses a fully Bayes approach to find the significant variables for the several specific population in the data set. I’ll use the significant variable found by him to construct a causal graph and also the change of patient’s characteristic given the manipulation variable, i.e., the comparisons.

**Case vs control patients**

These patients are divide into 2 specific population, used as the manipulation variable, which are: case and control. The other characteristics of those patient aren’t manipulated whatsoever. Ang Li’s full Bayes method concluded that the following variables are the significant ones for these division:

1. wa = African ancestry variable

2. rs7824364 (R1) = SNP Genetic ancestry and prostate cancer susceptibility in PR and AA men
Figure 25: Optimal causal graphs with their respective p-value: a) case patients; b) control patients
3. rs7210100 (R2) = prostate cancer SNP in AA

4. cigarettes smoker (CS) = current, previous and non-smoker with values 1, 2, 3 respectively

**Background knowledge**

We know that the first characteristic develop by the subject we are studying is his genetic, so the variables which are considered significant for this specific population cannot cause “wa”. Next, it follows the chromosomes associated with the prostate cancer and finally if the person is a cigarette smoker. In all of the cases, I will try to find a causal graph between those cases and find if there is a change in characteristics of the subjects and its causal structure. Using the background knowledge and TETRAD program to find the association using the Markov conditions and we got the following optimal causal graph with his p-value respectively for the control population:

As you can see, the p-value is high so we can consider it as a optimal causal graph, now the multi-linear regression equations are as following but first:

\[ CS = \epsilon_{CS} \]

\[ wa = \epsilon_{wa} \]

\[ R2 = 0.2932wa - 0.0325 + \epsilon_{R2} \]

\[ R1 = 1.0541wa + 0.1468\epsilon_{R1} \]

Since \( 0 < wa < 1 \), then “wa” doesn’t have any influence on “R2”, i.e. even it has a continuous associa-
tion, it doesn’t change R² since it’s a discrete variable (values are, 0,1 or 2). In terms of statistics, the difference \( P(R²|wa) - P(R²) = 0 \) for all the values of R² and wa, therefore the second condition of prima facie doesn’t satisfy. However, for R1, for at least one allele be present in the subject, the characteristic “wa” has to be greater than 0.81, i.e., the subject need to have 81% of African Ancestry in his genetics. So, \( |P(R1|wa) - P(R1)| > 0 \) for \( wa > 0.95 \) as for \( wa < 0.95, \ P(R1|wa) - P(R1)| = 0.\)

Now for the case population using the same background knowledge and program we have the following optimal graph and the p-value respectively:

In this case we have a new connection, “wa” variable causing “CS”. Since the p-value is low, then this optimal causal graph isn’t good enough to make conclusion. However, the connection seems interesting, it implies that his genetics influence on the his smoking behavior, but as you know, \( 0 < wa < 1 \), the regression equation is as follows:

\[
CS = 0.5982wa + 1.7954 + \epsilon_{CS}
\]

and again, CS is a categorical distribution (0,1,2), hence \( P(CS|wa) - P(CS) = 0.\)

**Stage of tumor >pT2 and ≤ pT2**

In this experiment, the population is divided in patients which the interested characteristic division is the tumor stage. Ang li’s full bayes method found that the following variables are significant in the data set:

1. gleason score (GS) = gleason1 + gleason2
2. na = native ancestry
3. Pre control serum (PSA) = in ng/ml

4. surgicalmarginstatus (SMS) = values: positive (1), negative (0)

**Background knowledge**

Again, the genetics cannot be an effect. The PSA levels in men doesn’t mean that they have prostate cancer (shown in recent studies), since having high PSA levels have two definitions for the patient: having prostate cancer or inflammation of prostate, so the gleason score doesn’t indicate that PSA are high neither PSA levels doesn’t indicate the gleason score, so the two possibilities that exist between the connection of those variables can be:

1. they are independent. No association at all.
2. they have a hidden common cause which is not considered in this experiment.

Now, there can be a higher chance that if the patient have prostate cancer, the PSA levels are high. Now, since “GS” and “PSA” variables could have a hidden common cause, they’ll be consider in the same level of causal relationship, i.e., we don’t know which is prior to the other one. Using the TETRAD program I got the following optimal causal graph for the population that have the stage level below 2 and his p-value respectively:

![Diagram](image)

Not a bad optimal causal graph, but analyzing the graph we have the following regression equations (I will not include the error terms):

\[ PSA = 0.9126GS - 0.3006 \]

\[ GS = -1.5484na + 6.8844 \]
Since $0 < na < 1$, the gleason score varies from 5 to 6. Now, my interest is the connection between “PSA” and “na” so using the technique of regression equation in causal graph we have that:

\[
\begin{align*}
&\Rightarrow PSA = 0.9126GS - 0.3006 \\
&\Rightarrow PSA = 0.9126(-1.5484na + 6.8844) - 0.3006 \\
&\Rightarrow PSA = -1.41307na + 5.9821
\end{align*}
\]

Again, $0 < na < 1$ then the PSA levels ranges from $4 < PSA < 6$ but in this experiment the PSA levels ranges from $0 < PSA < 74$ which is a significantly difference since there are only 120 patients of 359 that are in this range. So we can tell that “na” is a weak causal variable of “PSA” or it doesn’t influence much in “PSA”. This is the reason why the p-value is not high enough for this causal graph. In addition, there are many missing values that can be helpful in the experiment.

Now for the population with stage tumor level above 2, using the same Background knowledge we have the following optimal causal graph and p-value respectively:

What we can analyze between the two population is the change in characteristics between the patients. Notice that “na” isn’t associate at all with “GS”, neither “GS” with “PSA”. The connection between “PSA” and “SMS” have the following regression equation:

\[
SMS = 0.0134PSA + 0.1741
\]

Since $0 < PSA < 78$, the following equation tells us that for the patients who have PSA levels greater than 62 ng/mL (rounding 0.9 or more to 1 for SMS) have cancer cells in the tissue removed, but as for the
patients with stage tumor greater than 2, the probability that their PSA levels are below 55 and the surgical margin status is positive is 97%, this is due to a patient who has 78 ng/mL PSA level. If we eliminate that outlier value, using the PC algorithm, we have that:

Here you can see that there is no connection between the significant variables in this experiment, the reason is that there are many missing values and little characteristics included in this data set, so the only conclusion we can make is, comparing to the population of cancer tumor stage less than 2, that the genetics doesn’t influence in the size of the tumor, maybe as the cancer stage is bigger or equal to two, their growth isn’t a linear growth which can be connected to the assume prior variables, could be a random growth or a type of growth which can be deduced with a hidden cause (characteristic). We need more data and characteristics include in the data sets so we can make clear conclusion about the connection between the variables.
7 Conclusion

Since there are common causes, hidden variables or characteristics of the subject we are studying which aren’t included in the experiment, I can’t construct a good causal relationship with the interested variables with the given data set. My recommendation is to try to get all characteristic of that subject which has a connection with the topic we are studying. Causality is a topic that need deep observation between the variables, any slightly change has to be considered in causal study since a bit change can tells us about the status of it. If you notice that if we study a quantity of people, every person has a distinct value to the other (most of the time) this is because there exist a hidden variable which can tells us why they have that certain value. The deeper we get to the information of the subject, the better the causal relationship we can get.
Appendix: Graphs definition summary

Graph = is a ordered triple \( < V, M, E > \), where

\[
V = \text{set of vertices} \\
M = \text{set of marks of the end of edges, i.e., EM(\text{empty mark}), o, >} \\
E = \text{order pair, edges that connect the vertices with marks, i.e., } [V_1, M_1], [V_2, M_2], \ldots
\]

If there is a directed edge from \( V_1 \) to \( V_2 \), then \( V_1 \) is a parent of \( V_2 \), and \( V_2 \) is a child of \( V_1 \)

Parents(\( V \)) = \text{set of all parents of vertices in } V \\
Children(\( V \)) = \text{set of all children } \in V \\
\text{Indegree} = \text{number of vertex } V \text{ parents} \\
\text{Outdegree} = \text{number of vertex } V \text{ children} \\
\text{Degree} = \text{number of vertices adjacent to } V \text{ (Indegree + outdegree in directed graph)}

Undirected path between \( V_1 \) and \( V_2 \) \( \in G \) = sequence of vertices beginning with \( V_1 \), and ending \( \in V_2 \), i.e.,

\[ [V_1, M_1], [X_1, M_1], [V_1, M_1], [V_2, M_2], \ldots, [V_n, M_n], [V_2, M_2] \]

An edge \( [X, M_1], [Y, M_2] \) is in the path \( U \) iff \( X \) and \( Y \) are adjacent to each other in \( U \)

If the edge containing \( X \) in a undirected path between \( X \) and \( Y \) is out of \( X \) then we say that the path is out of \( X \); If \( X \) is into, then the path is into \( X \)

Empty path = sequence that consist of a single vertex

Acyclic = a path that contains no vertex more than once; otherwise cyclic

Two path intersects iff they have a vertex in common

If path \( U = < U_1, \ldots, U_n >, V = < U_n, \ldots, V_m > \), concatenation = \( < U_1, \ldots, U_n, V_i, \ldots, V_m > \)

Directed path between \( V_1 \) and \( V_2 \) \( \in G \) = sequence of vertices beginning with \( V_1 \) and ending \( V_2 \), and in the form \( [V_1, EM], [V_{(i+1)}, >] \).

For a directed edge \( e \) from \( U \rightarrow V \), head(\( e \)) = \( V \), and tail(\( e \)) = \( U \)

Directed acyclic graph = directed graph that contains no directed cyclic path

Semi directed path between \( A \) and \( B \) = undirected path that doesn’t contain an arrowhead pointing towards \( A \)

A graph is complete if every pair of vertex in \( G \) is adjacent

A graph is connected if there is a undirected path between any points

A subgraph of \( < V, M, E > \) is any graph \( < V', M', E' > \) such as \( V' \in V, M' \in M, E' \in E \)

Clique = subgraph of \( G \) that is complete.
Triangle in \( G \) = complete subgraph of \( G \) with three vertices.

Ancestor of vertex \( V \) = any vertex \( W \) such that there is a directed path from \( W \) to \( V \).

Descendent of \( V \) = \( W \in G \), such that there is a directed path from \( V \) to \( W \).
9 References


