# *Understanding statistical paradoxes using causal graphical models*

## *STA4517 Final report*

Name: Tian Han **Guan**

Student ID: 998978058

Department: Statistics

Program: MSc

Email: tianhan.guan@mail.utoronto.ca

---

**Abstract**

At the root of scientific explanations, causal inference plays a significant role in many fields such as social studies, economics, and artificial intelligence. However, making causal conclusions in many cases can be very challenging as in general causation is not the same as association and making decisions solely based on statistical methods can be counterintuitive and misleading. This report summarizes a few famous statistical paradoxes and discusses how causal inference techniques can be used to explain those. Among the many causal inference frameworks, the report focuses on the causal graphical model which is widely used in econometrics and artificial intelligence. Both the backdoor and frontdoor adjustment formulas are discussed as how they can be applied to resolve the paradoxes through simulated observational data examples.

---

# 1. Introduction of causal inference

Compared to statistical inference which uses data analysis methods to measure the association relationships between variables, causal inference is the process of making conclusions about the causal relationships between variables. In other words, causal inference is used to decide whether changes in one variable will cause direct changes on another. In general, association does not imply causation. For instance, causation, reverse causation, and confounding can all cause an association relationship to be seen, and therefore, causation is not always the only source of association and equating association to causation can be misleading in making conclusions about causality. As a result, knowledge of causal inference becomes essential in almost all scientific and social explanations.

There are three commonly used frameworks for causal inference, namely, the potential outcome framework, the structural equations model framework, and the graphical model framework. Although all three frameworks share many common characteristics, they are usually used in different fields for different purposes. This report focuses on the framework of graphical models (also known as Bayesian networks), and discusses how it can be used to understand various statistical paradoxes that were recognized by people historically. One major advantage of using graphical models for causal inference is that they can provide a clear representation of the independence structure of the probabilistic model visually, and therefore making the model more accessible to non-mathematical audience. In addition, many of the numerically intense inference and learning tasks about probabilistic models can be solved easily using graphical models, making more problems approachable than traditional methods. As a result, graphical models are widely used in many fields such as biostatistics, econometrics, and artificial intelligence.

# 2. Causal graphical models

## 2.1 Basics of graphical models

As one of the most commonly used technique for many statistical and machine learning problems, graphical models (or more specifically, probabilistic graphical models) had drawn numerous attentions in the past few decades due to their ability to provide a compact and clear representation of joint probability distributions and conditional independence structure between random variables. A graphical model consists of two components: nodes and edges, where nodes represent random variables and edges represent the dependence relationship between variables. There are two types of probabilistic graphical models, undirected graphical models and directed graphical models. Although both types are used to encode the factorization and dependence information, they are completely different in terms of how they encompass dependence relationships and induced factorization properties. Undirected graphical models, also known as *Markov Random Fields (MRF) or Markov Networks*, represent the conditional independence of two nodes given a third if all paths between the two nodes are separated by the third node. An undirected graphical model can be either acyclic or cyclic (and can therefore encompass cyclic dependence structures). On the other hand, directed graphical models, also known as *Bayesian Networks*, are acyclic and have directions on the edges, can encode much more complicated dependence structure such as the ones that are induced by the causal relationship between random variables. This report focuses on the directed graphical models (or Bayesian Networks) as they are the ones that are widely used in causal inference to understand the causal relationships between variables.

## 2.2 Bayesian Network

A Bayesian network (also known as belief network or probabilistic directed acyclic graphical model) is a type of graphical model that is used to represent the dependence structure of a set of random variables using a directed acyclic graph (*DAG*). The following is a formal definition of a Bayesian network:

**Definition 1 (Bayesian network)**

A Bayesian network is a directed graph $G(V, E)$ with the following:

1) A random variable $X_i$ for each node $i \; in \; V$
2) Satisfies the local Markov property (or *Markov blanket property*) as explained later

In summary, a Bayesian network defines a joint distribution $p(x_1, x_2, \dots, x_n)$ for $(X_1, X_2, .., X_n)$ . A joint distribution $p(x_1, x_2, \dots, x_n)$ can be factorized over a Bayesian network G based on a product of factors specified by G.

**Remarks**

1) It can be shown that a joint distribution $p(x_1, x_2, \dots, x_n)$ represented by a Bayesian network must be a valid *probability mass function (pmf)*, in other words, $p(x_1, x_2, \dots, x_n)$ must be non-negative over all possible configurations in the state space of $(X_1, X_2, .., X_n)$ and sum to one over all possible configurations.
2) It can be shown that a Bayesian network must be acyclic: if G contains any cycles, then the represented joint distribution $p(x_1, x_2, \dots, x_n)$ may not be a valid pmf (i.e. non-negative everywhere and sums to one).

## 2.2.1 Factorization of a Bayesian Network

**Definition 2 (Parent and Child)**

Node X is said to be a ***parent*** of node Y (and Y is called a ***child*** of X) if there exists a ***directed edge*** from X to Y. Note that a node can have more than one parent and more than one child. Denote the set of parents of node Y by ***pa(Y)***.

**Definition 3 (Ancestor and Descendant)**

Node Z is said to be an ***ancestor*** of node W (and W is called a ***descendant*** of Z) if there exists a ***directed path*** from Z to W. Similarly, a node can have more than one ancestor and more than one descendant. Denote the set of ancestors of node W by ***an(W)*** and the set of descendants of node Z by ***de(Z)***. By definition, a node is always the ancestor and descendant of itself.

**Definition 4 (Local Markov Property)**

Denote X to be all the nodes in a graph G, denote $X_\alpha$ to be some node in G and $N[X_\alpha]$ to be the set of nodes in the *neighborhood* of $X_\alpha$, then the Local Markov Property (also known as *Markov Blanket Property*) says that:

$$\boldsymbol{X_\alpha} \perp \mathbf{X} \backslash (\boldsymbol{N}[\boldsymbol{X_\alpha}] \cup \boldsymbol{X_\alpha}) \mid \boldsymbol{N}[\boldsymbol{X_\alpha}]$$

The property says that a variable is conditionally independent of all other variables (except its neighbors and itself) given its neighbors. Intuitively, this implies that given all the variables

adjacent to a variable X, then all non-adjacent variables of X will provide *no useful information* in learning about the distribution of X.

**Derive the factorization**

Bayesian networks always satisfy the above local Markov property, which implies that a variable is always conditionally independent of its non-descendants variables given its parents. The local Markov property allows us to obtain a much simpler factorization of the joint distribution with the Bayesian network:

By the total law of probability, we know that any joint distribution can be factorized as follows:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

$$= P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1) \dots P(X_n = x_n|X_{n-1} = x_{n-1}, \dots, X_1 = x_1)$$

$$\equiv p(x_1)p(x_2|x_1) \dots . p(x_n|x_{n-1}, x_{n-2}, \dots, x_1)$$

The using the local Markov property, the above factorization can be simplified as:

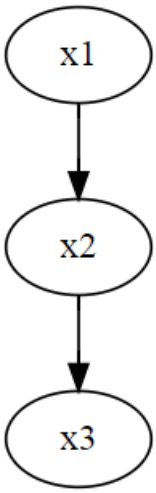
$$= p(x_1|pa(x_1))p(x_2|pa(x_2)) \dots . p(x_n|pa(x_n))$$

$$= \prod_{k=1}^{n} \boldsymbol{p(x_k|pa(x_k))}$$

### 2.2.2 Fundamental structures in a Bayesian network

In order to obtain the conditional independence information from a Bayesian network, it is essential to understand the three possible types of fundamental structures for every path of three variables: a chain, a fork, or a collider.

**Type 1: Chain (or Cascade)**



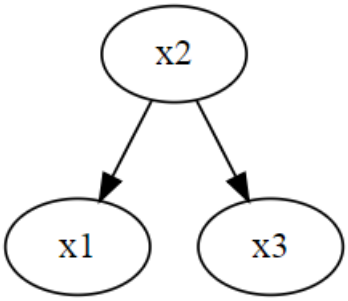

$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$

$since\ X_1\ has\ no\ parents, pa(X_2) = \{X_1\}, pa(X_3) = \{X_2\}$

Conditional independence in a chain:

$$\boldsymbol{X_1 \perp X_3 \mid X_2}$$

**Type 2: Fork (or Common parent)**



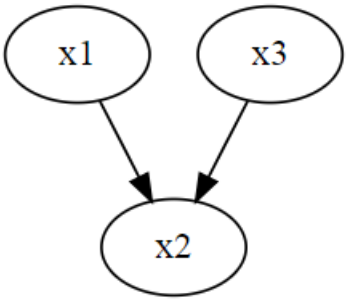

$p(x_1, x_2, x_3) = p(x_2)p(x_1|x_2)p(x_3|x_2)$

$since\ X_2\ has\ no\ parents, pa(X_1) = \{X_2\}, pa(X_3) = \{X_2\}$

Conditional independence in a fork:

$$\boldsymbol{X_1 \perp X_3 \mid X_2}$$

**Type 3: Collider (or V-structure/Explaining away)**




$p(x_1, x_2, x_3) = p(x_1)p(x_3)p(x_2|x_1, x_3)$

$since\ X_1\ and\ X_3\ have\ no\ parents, and\ pa(X_2) = \{X_1, X_3\}$

Note that in a collider, we only have independence structure $\boldsymbol{X_1 \perp X_3}$, but not the conditional independence $\boldsymbol{X_1 \perp X_3 \mid X_2}$, and therefore two independent variables will become dependent if the third variable is conditioned on.

The above three types of fundamental structures can represent the dependence structure of any three-variable Bayesian network. A general Bayesian network can be obtained by repeatedly apply the three fundamental structures over any bigger directed graph.

## 2.2.3 Remarks about Bayesian networks

**Remark #1: Markov Equivalence**

As mentioned earlier by its definition, we can always obtain a factorization of the joint distribution $p(x_1, x_2, \dots, x_n)$ from a Bayesian network, but it should be noted that given a factorization, one can possibly construct more than one Bayesian network which can all represent the given factorization correctly. This can be explained as follows:

- There can be more than one set of marginal joint distributions that all imply the same total joint distribution and conditional distributions (which are given in the factorization).
- Therefore, when the given factorization is only a special form of the most general factorization of the distribution, all structures which can generate the given conditional distribution is a valid Bayesian network for the given factorization.
- Two Bayesian networks are called ***Markov equivalent*** if they imply the same set of conditional independence structures (and therefore the same distribution)

**Theorem 1 (Markov equivalence)**

Two Bayesian networks are said to be ***Markov equivalent*** if and only if they have the same skeletons (or same undirected structures) and the same set of unshielded colliders (or v-structures).

As a result, Markov equivalence will give us difficulties to infer causal relationships from the probability distributions alone, in other words, we cannot distinguish Bayesian networks that are Markov equivalent without additional information about the marginal joint distributions.

**Remark #2: Markov blanket**

**Definition 5 (Markov blanket)**

The Markov blanket of a variable X, denoted by ***Mb(X)***, is the *smallest set* containing all variables carrying information about X that cannot be obtained from any other variable. In other words, the Markov blanket contains all the variables that shield the given variable from the rests of the variables. Markov blanket of a variable is the only set of variables that can provide useful information about that variable.

In a Bayesian network, the Markov blanket of a variable X is the set that consists of

1) Parents of X
2) Children of X
3) Parents of children of X

It can be shown that a variable X is conditionally independent of all other variables given its Markov blanket, in other words, for any variable Y that is not in the Markov blanket Mb(X), we have:

$$P(X|Mb(X), Y) = P(X|Mb(X))$$

## 2.2.4 Conditional independence and d-separation

One important feature of a Bayesian network is that it allows us to read off all the conditional independence relationships from the graph through the concept of ***d-separation***.

**Definition 6 (d-separation)**

Let π denotes a path between two variables X and Y (consists of adjacent variables). Then π is said to be *d-separated* by a set of variables Z if any of the following conditions holds true:

- π contains a chain, such that the middle variable W of the chain is in Z
- π contains a fork, such that the middle variable W of the fork is in Z
- π contains a collider (or V-structure), such that the middle variable of the collider W is NOT in Z and no descendants of W is in Z

Two variables X and Y are said to be d-separated by the set of variables Z if all paths between them are d-separated by Z.

**Theorem 2 (d-separation and conditional independence)**

Let $G(V, E)$ denotes a Bayesian network with set of variables V and let P denotes any distribution which can be factorized according to G. Let $X, Y$ be two variables in V. Then we have the following:

**X and Y are d-separated by Z => $X \perp Y \mid Z$**

In other words, d-separation implies conditional independence.

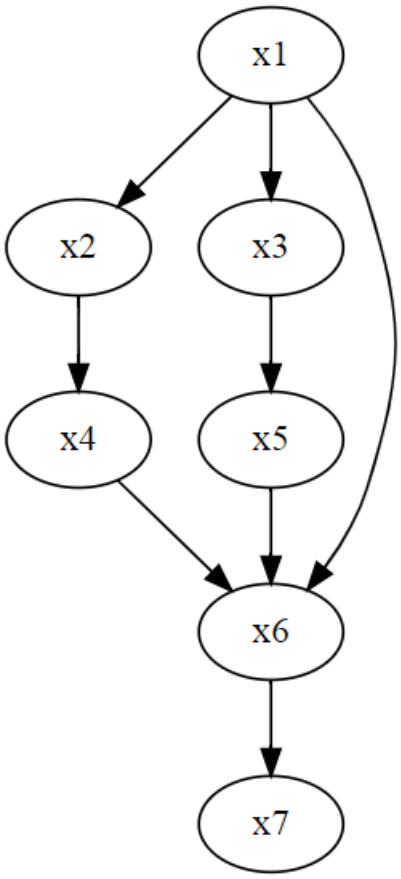
**Remarks**

1) If two variables X and Y are NOT d-separated, they are called d-connected.
2) Although d-separation implies conditional independence, d-connection does NOT always guarantee conditional dependence (even though under most of the cases this is true).
3) The converse of Theorem 2 is NOT true, i.e. conditional independence does NOT always guarantee d-separation as there can exist additional independence relationships.
4) Remarks 2 and 3 are true under the faithfulness assumption (assume no additional independence structure).
5) Intuition about d-separation and conditional independence:

   Given two variables X and Y, there exists a set of paths that connects X and Y. A path can be thought as *active* if it contains dependence relationship between X and Y given some set Z. Therefore, X and Y are d-connected given Z if at least one of the paths contain dependence information, otherwise, X and Y are d-separated given Z.

The following example illustrates the concept of d-separation.

**Example 1 d-separation**




Using the above Bayesian network, we want to answer the following 4 questions:

1) Is $X_1 \perp X_4 \mid X_2$?
2) Is $X_1 \perp X_4 \mid (X_2, X_6)$?
3) Is $X_1 \perp X_4 \mid (X_2, X_7)$?
4) Is $X_1 \perp X_4$?

From the graph, there are *three* different paths from $X_1$ to $X_4$:

$Path\ 1: X_1 \rightarrow X_2 \rightarrow X_4$

$Path\ 2: X_1 \rightarrow X_3 \rightarrow X_5 \rightarrow X_6 \leftarrow X_4$

$Path\ 3: X_1 \rightarrow X_6 \leftarrow X_4$

1) Conditioning on $Z = X_2$

- Path 1 is a chain with middle variable $X_2$. Since $X_2 = Z$, $X_1$ and $X_4$ are d-separated.
- Path 2 contains the collider $X_5 \rightarrow X_6 \leftarrow X_4$ with middle variable $X_6$. Since $X_6$ $is\ NOT\ in\ Z$ and all the descendants of $X_6$ ($here\ is\ just\ X_7$) are not in Z, $X_1$ and $X_4$ are d-separated.
- Path 3 is a collider again with middle variable $X_6$. Therefore, $X_1$ and $X_4$ are d-separated.

Since $X_1$ and $X_4$ are d-separated by $Z = X_2$ for all paths between them, $X_1$ and $X_4$ are d-separated by $X_2$ and therefore $\boldsymbol{X_1 \perp X_4 \mid X_2}$ **is true**.

2) Conditioning on $Z = (X_2, X_6)$

- Path 1 is a chain with middle variable $X_2$. Since $X_2$ $is\ in\ Z$, $X_1$ and $X_4$ are d-separated.
- Path 2 contains the collider $X_5 \rightarrow X_6 \leftarrow X_4$ with middle variable $X_6$. Since $X_6$ $is\ in\ Z$, $X_1$ and $X_4$ are NOT d-separated.

Since $X_1$ and $X_4$ are NOT d-separated by $Z = (X_2, X_6)$ for all paths between them, $\boldsymbol{X_1 \perp X_4 \mid (X_2, X_6)}$ **is false**.

3) Conditioning on $Z = (X_2, X_7)$

- Path 1 is a chain with middle variable $X_2$. Since $X_2$ $is\ in\ Z$, $X_1$ and $X_4$ are d-separated.
- Path 2 contains the collider $X_5 \rightarrow X_6 \leftarrow X_4$ with middle variable $X_6$.
  Although $X_6$ $is\ NOT\ in\ Z$, the descendant of $X_6, X_7$ is in Z, and therefore $X_1$ and $X_4$ are NOT d-separated.

Since $X_1$ and $X_4$ are NOT d-separated by $Z = (X_2, X_7)$ for all paths between them, $\boldsymbol{X_1 \perp X_4 \mid (X_2, X_7)}$ **is false**.

4) Unconditional independence (here Z can be thought as the empty set)

- Path 1 is a chain with middle variable $X_2$. Since $X_2$ $is\ NOT\ in\ Z$, $X_1$ and $X_4$ are NOT d-separated.

Since $X_1$ and $X_4$ are NOT d-separated by $Z = empty\ set$ for all paths between them, $X_1$ and $X_4$ are NOT d-separated and therefore $\boldsymbol{X_1 \perp X_4}$ **is false**.

One important observation from this simply example is that conditioning on "extraneous" confounding variables can actually kill the conditional independence between two variables. This will be illustrated in detail in later sections using the famous Simpson's paradox.

## 2.3 Causal Bayesian networks

### 2.3.1 Basics of causal Bayesian networks

Although Bayesian networks are very useful in terms of determining the conditional independence structures of a set of variables, in many cases, they may not give the correct causal relationships. For example, both a chain and fork imply the same conditional independence relationship $\boldsymbol{X_1 \perp X_3 \mid X_2}$, but obviously they have completely different causal relationships. As

a result, causal Bayesian networks are developed to make sure that the relationships in a Bayesian network are actually causal using some additional requirements.

**Definition 7 (Causal Bayesian network)**

Denote $P(v)$ be a probability distribution on a set of variables V, and let $P_x(v)\ or\ equivalently\ P(v|do(x))$ denote the distribution resulting from the intervention $do(X = x)$ (i.e. sets the set of variables $X \subseteq V$ to some constant value x). Then let $P_*$ denote the interventional space (i.e. set of all interventional distributions $P_x(v)\ , X \subseteq V$, including $P(v)$ itself (which means no intervention).

Then a Bayesian network G is called a *causal Bayesian network* compatible with $P_*$ if and only if the following conditions hold true for every $P_x(v)\ \subseteq P_*$:

1. $P_x(v)$ is Markov relative to the Bayesian network G
2. $P_x(v_i)\ =\ 1$ for all $V_i\ \epsilon\ X$ whenever $v_i$ is consistent with $X = x$
3. $P_x\left(v_i\middle|pa(X_i)\right) = P(v_i|pa(X_i))$ for all $V_i\ \notin X$ whenever $PA(X_i)$ is consistent with $X = x$

## 2.3.2 Causal assumption and invariance assumption

The definition of the causal Bayesian network implies the following *causal assumption*:

- Given the parents of a variable $X_i$, $X_i$ can be represented by the following function:

  $\boldsymbol{x_i = f_i(pa(x_i), \epsilon_i)}, where\ \epsilon_i's\ are\ mutually\ independent\ random\ noise$

This essentially means that in a causal Bayesian network, if a variable X is caused to be in a state x, denote the action to be do($X = x$), then the probability density function is obtained by removing all the edges from *pa(X) to X* and set the value of X to be x. It should be also noted that, under the causal assumption, when we make intervention on one variable, the structure of the causal graph and the functional relationships between the other variables will always remain the same (i.e., we have the *invariance assumption*).

As a result, the factorization of a causal Bayesian network is the factorization of a Bayesian network plus the additional causal assumption that $x_i = f_i(pa(x_i), \epsilon_i)$, i.e.

$$\boldsymbol{P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{k=1}^{n} p\left(x_k \middle| pa(x_k)\right), x_i = f_i(pa(x_i), \epsilon_i) \; for \; i = 1, .., n}$$

**<u>Some comments on causal graphical models</u>**

1) Compared to probabilistic graphical models which are designed to encode the statistical relationships between variables, causal graphical models are designed to make inference and predictions on the effect of interventions or actions. By adding the feature of cutting variables off from their parents given an intervention, the effect of interventions can be obtained and estimated.
2) It should be noted that causal graphs can only tell us if there exists a causal relationship between variables through conditional independence, but they cannot tell us what the relationship between variables are (e.g. how strong the causal relationship is).

### 2.3.3. Do-calculus and the do-operator

Do-calculus (also known as Peral's causal calculus or calculus of actions), was first developed by J. Pearl in 1995 to facilitate the identification of causal effects in graphical models.

A new operator do() operator was developed to mathematically define an intervention by removing all the edges that are going into the target of intervention (i.e. removing all the probability factors in the factorization of the intervention) while keeping everything else the same by setting the variable to some constant value. The causal effect of X on Y is denoted by $\boldsymbol{P(y|do(X=x))}$ $\boldsymbol{or\ simply\ P(y|do(x))}$, which represents an interventional distribution and is distinguished from the observational distribution $\boldsymbol{P(y|x)}$. The ultimate goal of do-calculus is to be able to determine the interventional distribution in terms of the observable distribution so that the effect of intervention can be determined.

### 2.3.4. Confounding bias

One major problem in causal inference is whether we should control the confounders when we assess the causal effect of one variable X on another variable Y, and if so, which of the confounders should we control. For example, it is a common approach for people to partition the whole population/sample into strata according to the confounders (for example, when evaluate how effective a new treatment is, often divide the population/sample into homogenous strata based on factors such as gender, age, etc) and then evaluate the treatment effect in each stratum and then report the average treatment effect across all strata. However, in many cases, people realize that the statistical relationship between variables can be reversed by including different

confounders in the statistical analysis, and this phenomenon is recognized as the famous *Simpson's paradox*. The following section summarizes how to adjust for confounders using causal graphical models.

## 2.3.5. Backdoor adjustment formula

Given a causal Bayesian network $G(V, E)$ with observational data on the set of variables V, we are usually interested in estimating the effect of some intervention $do(X = x)$ on some sets of variables Y, in other words, $P\left(y \middle| do(X = x)\right)$**.** This can be done using the famous ***backdoor adjustment formula***, which states that under certain conditions, how we can estimate the causal effect of X on Y given a set of variables $W \subseteq V$ .

**Definition 8 (Backdoor criterion)**

A set of variables $W \subseteq V$ satisfies the backdoor criterion relative to variables $(X, Y)$ in a Bayesian network $G(V, E)$ if W satisfies the following conditions:

- W blocks **all backdoor paths** between X and Y (a backdoor path is a path that contains an directed edge or arrow into X)
- W does not include any **descendants of X** (which means W does not contain X itself by definition)

**Comments:**

1) The parents of X **pa(X)** always satisfies the backdoor criterion

2) The backdoor criterion can be generalized into the condition where X and Y are two disjoint sets of nodes, in this case, the backdoor criterion is satisfied if the criterion is satisfied for every pair of variables $(X_i, Y_j)$ $such\ that\ X_i \epsilon\ X, Y_j\ \epsilon\ Y$

**Theorem 3 (Backdoor adjustment formula)**

Given a set of variables W which satisfies the backdoor criterion relative to variables $(X, Y)$, the causal effect of X on Y is identifiable and is given by the following formula:

$$\boldsymbol{P(y|do(x)) = \sum_{w} P(y|x,w)P(w)}$$

**<u>Some remarks on the backdoor adjustment formula</u>**

1) Intuitively, the backdoor adjustment formula shows that blocking (or conditioning on) all backdoor paths will adjust for all biases induced by the confounding variables. In addition, excluding all descendants will prevent new paths to be created which could invoke additional new bias.
2) The formula also tells us that if there exists confounding variables (which create the backdoor paths), then the observed association between variables X and Y can either be caused by a direct causal relationship between X and Y or can be caused by some or all of the confounding variables (the backdoor paths). Again, this shows that association is not always the same as causation.
3) In order to use the backdoor adjustment formula, we assume that sufficient number of confounding variables are being observed so that we can adjust all the backdoor paths. However, there can be scenarios where not enough number of confounding variables are

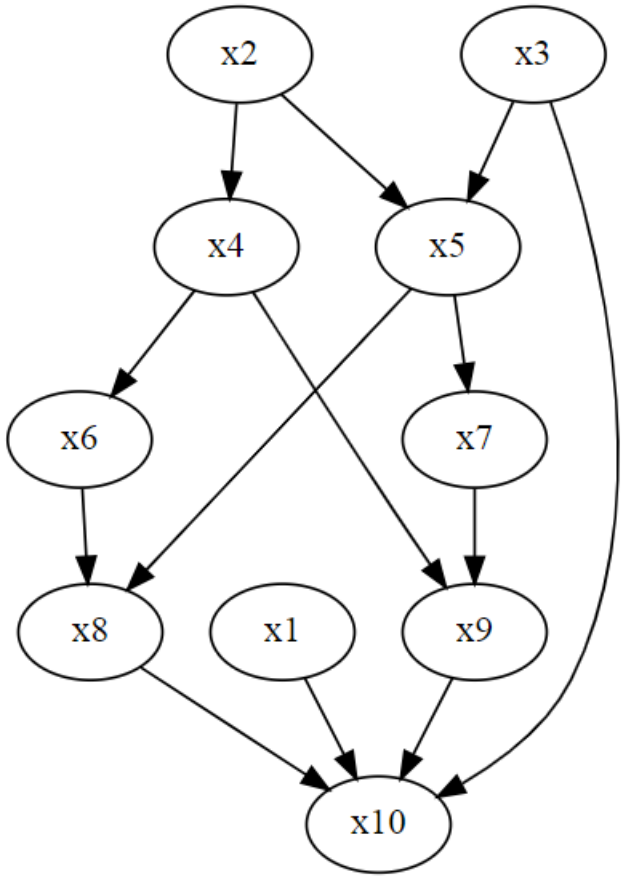
observed and therefore we cannot adjust for all backdoor paths. In such scenarios, the frontdoor adjustment formula can be used to evaluate the causal inference.

4) If there are only directed paths between the two variables X and Y (this means that there are no confounding variables for X,Y), then the causal effect can be easily estimated by the association effect, in other words, $P(y|do(x)) = P(y|x)$
5) From a graphical model perspective, confounding variables W can also be viewed as ancestor variables that both X and Y share in common. Thus, $P(y|do(x)) \neq P(y|x)$ and estimate $P(y|do(x))$ using $P(y|x)$ will cause the estimates to have confounding bias.

**Example 2 Backdoor criterion**

This simple example illustrates the concept of backdoor criterion. Consider the following causal graph and we are interested in the following two questions:

1. Find all the backdoor paths from $X_5$ to $X_9$
2. Find all sets of variables that satisfy the backdoor criterion with respect to the variables $X_5$ and $X_9$




1) All the backdoor paths from $X_5$ to $X_9$ are listed below:

$Path$ 1: $X_5 \leftarrow X_2 \rightarrow X_4 \rightarrow X_9$, the middle variables can be classified as:

- $X_2$ acts a fork, $X_4$ acts as a chain
- How to block this path? [Block $X_2$ *or* $X_4$ *or both of them*]

$Path$ 2: $X_5 \leftarrow X_3 \rightarrow X_{10} \leftarrow X_9$ , the middle variables can be classified as:

- $X_3$ acts a fork, $X_{10}$ acts as a collider
- How to block this path? [No action (or equivalent to say DO NOT block $X_{10}$) as block on $X_{10}$ will open this backdoor path]

$Path\ 3$: $X_5 \leftarrow X_3 \rightarrow X_{10} \leftarrow X_8 \leftarrow X_6 \leftarrow X_4 \rightarrow X_9$ , the middle variables can be classified as:

- $X_3$ $and$ $X_4$ act as forks, $X_{10}$ acts as a collider, $X_8$ $and$ $X_6$ act as chains
- How to block this path? [No action (or equivalent to say DO NOT block $X_{10}$) as block on $X_{10}$ will open this backdoor path]

$Path\ 4$: $X_5 \leftarrow X_2 \rightarrow X_4 \rightarrow X_6 \rightarrow X_8 \rightarrow X_{10} \leftarrow X_9$ , the middle variables can be classified as:

- $X_2$ acts as a fork, $X_{10}$ acts as a collider, and $X_4, X_6,$ $and$ $X_8$ act as chains
- How to block this path? [No action (or equivalent to say DO NOT block $X_{10}$) as block on $X_{10}$ will open this backdoor path]

2) Now, for any set of variables W to satisfy the backdoor criterion, it must be able to block all backdoor paths and does not include any of the descendants of $X_5$ (the set of descendant of $X_5$ includes $X_7$, $X_8$ , $X_9$ $and$ $X_{10}$) and obviously W should not contain $X_5$ $and$ $X_9$ as W is the set of confounding variables for them. This left with the other 5 variables $\{X_1, X_2, X_3, X_4, X_6\}$

From the analysis above, we know that W must include at least one of $X_2$ $or$ $X_4$ in order to block path 1. In addition, since none of $X_1, X_3$ $or$ $X_5$ acts as a collider on any of the backdoor paths, including them in W will also be valid as they will not open any of the backdoor paths. In summary, a set of variables W satisfies the backdoor criterion for the above causal graph if satisfies both of the following two conditions:

1) W contains at least one of two variables $X_2$ $and$ $X_4$
2) W does not contain any of the variables in the set $\{X_5, X_7, X_8, X_9, X_{10}\}$

## Backdoor criterion and confounding bias

The most important contribution of the backdoor criterion is that it gives us a way of determining which set of confounding variables to include when we evaluate the causal effect of one variable on another. In other words, if a set of confounding variables satisfy the backdoor criterion, then the evaluation of the causal effect will be free of confounding bias, and on the other hand, if the set of confounding variables that are being adjusted does not satisfy the backdoor criterion, then the evaluated causal effect will not be trustable as it will contain some level of confounding bias. In the above example, the set $\{X_1, X_2, X_3, X_6\}$ satisfies the backdoor criterion and therefore using $\{X_1, X_2, X_3, X_6\}$ as confounding variables will adjust for all the confounding bias. On the other hand, the set $\{X_1, X_2, X_4, X_8\}$ does not satisfy the backdoor criterion and therefore using $\{X_1, X_2, X_4, X_8\}$ will introduce confounding bias in causal inference.

In summary, when adjusting for confounding effect, it is important to choose the correct set of confounders to use, and failure to do so will cause the interpretations to fail. This leads to the famous Simpson's paradox in the next section.

# 3. Understanding Simpson's paradox, birth-weight paradox, and Berkson's paradox

## 3.1 Introduction of Simpson's paradox

Simpson's paradox (also known as the Yule-Simpson effect, reversal paradox, or amalgamation paradox) is a statistical paradox in which the association between two variables disappears or reserves upon conditioning on a third variable, regardless of which value the third variable takes on. It is first mentioned by Edward H. Simpson in his technical paper in 1951, and statisticians Karl Pearson (1899) and Udny Yule (1903) had also mentioned the paradox in similar fashions.

The following example illustrates how Simpson's paradox can give counterintuitive results using traditional statistical reasoning and methods.

**Example 3 Simpson's paradox**

Suppose a marketing company has been using two types of advertising strategies over the past year, and now with the collected information (that is, we are dealing with observational data rather than randomized trials), it wants to assess which advertising strategy gives more increase in sales of the products which are running shoes. Here, advertising strategy #1 is internet and advertising strategy #2 is broadcast media (includes TV and radios). In the observational data, there are three variables: advertising strategy (denote it by X), increase in sales (denote it by Y), and gender (denote it by Z).

The marketing company decides to perform two types of statistical analyses:

Analysis 1: Assess which advertising strategy is better using the ***aggregated data*** (that is, just variables X and Y)

Analysis 2: Assess which advertising strategy is better using the ***segregated data*** (that is, using all three variables X, Y, Z, and see which advertising strategy is better in each of the two gender groups)

In practice, many people will perform these two types of analysis and expect them to give the same conclusion (usually people use both types so that one analysis can support the other and give the audience more information). However, as we will see below, in this simulated data example, the two types of analysis give completely different conclusions and therefore make the statistical analysis to be counterintuitive.

The following tables and graphs illustrate the results of the two types of analysis using the observation data.

A small sample of the observational data is shown as follows:

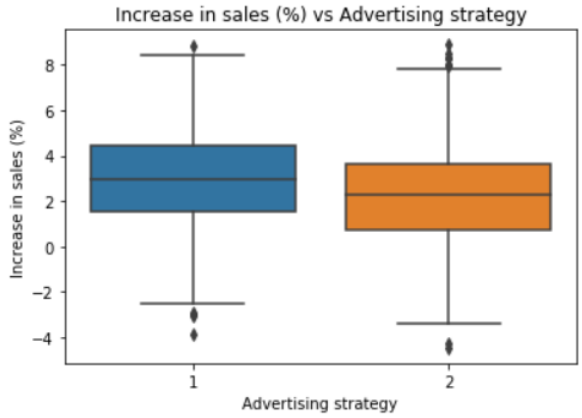
- Variable Z = Gender (0 = male and 1 = female)
- Variable X = Advertising strategy (1 = Internet, 2 = Broadcast)
- Variable Y = Increase in sales (in %)
- 2,000 samples are used in this example

| | Advertising strategy | Gender | Increase in sales (%) |
|---|---|---|---|
| **0** | 2 | 0 | -1.390892 |
| **1** | 1 | 1 | 4.324473 |
| **2** | 2 | 0 | 0.858401 |
| **3** | 1 | 1 | 4.775531 |
| **4** | 1 | 1 | 1.053707 |

| | Advertising strategy | Gender | Increase in sales (%) |
|---|---|---|---|
| **1995** | 1 | 0 | 0.639190 |
| **1996** | 1 | 1 | 2.883586 |
| **1997** | 1 | 1 | 4.345658 |
| **1998** | 1 | 1 | 2.913871 |
| **1999** | 1 | 1 | 1.971229 |

## Analysis 1: Results from the aggregated data

### Chart 1.1 Compare the results of both advertising strategies using boxplots




### Chart 1.2 Compare the results of both advertising strategies using violin plots

**Table 1. Compare the average increase in sales (in %) of both advertising strategies using the aggregated data**

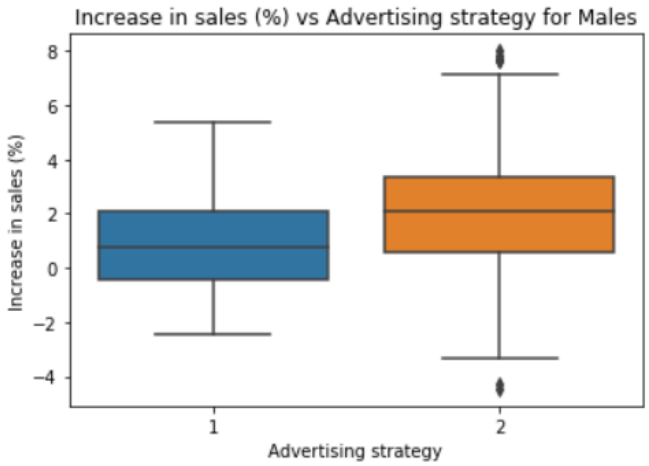
| Advertising strategy | Increase in sales (%) |
|---|---|
| **1** | 2.886823 |
| **2** | 2.224681 |

**Conclusion:**

From both the table summary and the visual plots, it is clear that ***advertising strategy 1 (internet advertisement)*** is the better strategy based on the observational data.

## Analysis 2: Results from the segregated data

**Chart 2.1 Compare the results of both advertising strategies using boxplots**



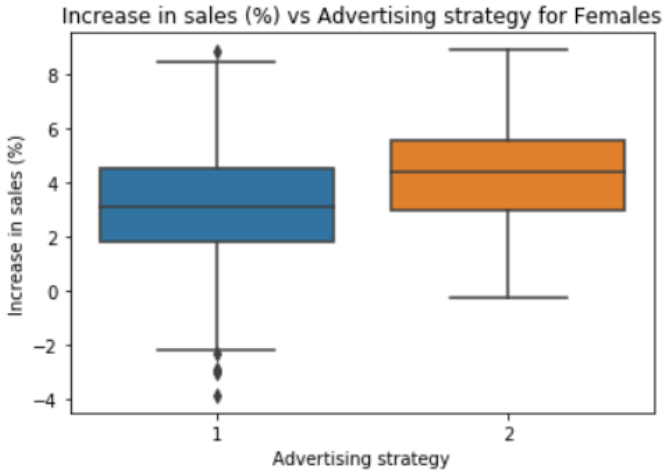



**Chart 2.2 Compare the results of both advertising strategies using violin plots**


**Table 2. Compare the average increase in sales (in %) of both advertising strategies using the segregated data**

| | | Increase in sales (%) |
|---|---|---|
| Gender | Advertising strategy | |
| 0 | 1 | 0.897125 |
| | 2 | 2.000512 |
| 1 | 1 | 3.087558 |
| | 2 | 4.235542 |

**Conclusion:**

From both the table summary and the visual plots, it is clear that ***advertising strategy 2 (broadcast advertisement)*** is the better strategy based on the observational data.

**What's going wrong here?**

This simple example illustrates the idea of Simpson's paradox: if we partition the population into strata based on certain characteristics (or a third variable), then the measured associations are completely different between the aggregated data and the segregated data (in this example, the sign of the association actually got *reversed*). Given both types of analyses are statistically correct, the question is which analysis should we use? It turns out that in order to resolve Simpson's paradox and make the correct decision, we will need the help of a causal graph and use the famous backdoor criterion.

## 3.2 Resolve Simpson's paradox using causal graphical models

### 3.2.1 The sure-thing theorem

Simpson' paradox has drawn many attentions in the past century and is labelled as a "paradox" due to its physical impossibility from a decision-making perspective. Here, it should be noted that although *reversals in association* when partition (or condition) the population on a third variable is possible, *reversals in causation* is never possible. For example, in the previous example, it is impossible that broadcast advertisement is the better advertisement strategy for both male and female customer groups but is the worse strategy for the whole customer population. This impossibility phenomenon is formally defined using the so-called sure-thing principle in decision theory.

**Theorem 4 (Sure-thing principle)**

A decision maker who would take a certain action if he knew that an event E has occurred or the negation of E has occurred, should also take the same action if no information of E is given.

In simple words, the sure-thing principle suggests that the uncertainty information of an event is useless in a decision making process if the event will not affect the decision making anyway. The principle tells that worrying about irrelevant uncertainties is a waste of effort.

Jeffrey and Pearl later on has shown that the sure-thing principle is only valid when the probability of the considered event E is irrelevant to the action. In this case, using Pearl's do-calculus, the sure-thing principle can be stated as the following sure-thing theorem.

**Theorem 5 (Sure-thing theorem, Pearl 2009)**

Given the fact that an action A does not change the distribution of the subpopulations (that is, A has the same treatment effect across all individuals in the population), then the causal effect of action A on event B must have the same direction on both the whole population and each subpopulations. In other words, the causal effect should not be reversed from whole population to each subpopulation. In addition, it should be noted that the sure-thing theorem holds true regardless of whether the third variable is a confounder or not (even though Simpson's paradox is caused by confounding bias).

## 3.2.2 Explain Simpson's paradox

From the sure-thing theorem (and assume the assumption in the theorem is true, that is, the action or treatment has the same causal effect for all individuals), then we know Simpson's paradox is impossible from a causal inference perspective. Indeed, Simpson's paradox is an example where people confuse between the concept of association and causation (or more generally, statistical inference and causal inference).

Consider the previous advertisement example again. From a causal inference perspective, we know the following three *causal statements* cannot happen at the same time:

1. Strategy #1 is better for male customers
2. Strategy #1 is better for female customers
3. Strategy #1 is worse for all customers

However, it is entirely possible that the following three *statistical or probabilistic statements* can happen at the same time:

1. $E[Increase\ in\ sales|Male, Strategy\ \#2] > E[Increase\ in\ sales|Male, Strategy\ \#1]$
2. $E[Increase\ in\ sales|Female, Strategy\ \#2] > E[Increase\ in\ sales|Female, Strategy\ \#1]$
3. $E[Increase\ in\ sales|\ Strategy\ \#2] < E[Increase\ in\ sales|Strategy\ \#1]$

As a summary, Simpson's paradox arises when people think the three *statistical or probabilistic statements* are the same as the three *causal statements* or more generally, association implies causation.

### 3.2.3 Making the right decision in Simpson's paradox

It turns out that in order to make the right decision in Simpson's paradox, we will need to use the backdoor criterion, that is, whether the third variable Z satisfies the backdoor criterion relative to the interested variables (X, Y). The following steps should be followed in a Simpson's paradox when association reversal is seen:

1. Identify whether the third variable Z satisfies the backdoor criterion relative to the variables (X, Y)
2. If Z satisfies the backdoor criterion relative to the variables (X,Y), then the segregated data should be used to make the decision by conditioning on Z. Otherwise, if Z does not satisfy the backdoor criterion relative to the variables (X,Y), then the aggregated data should be used to make the decision as conditioning on Z will cause confounding bias
3. It should be noted that there is no guarantee that the right decision lies in either the aggregated or the segregated data, that is, it is entirely possible that both datasets can give the

wrong answer. This happens when Z is not sufficient to block a backdoor, in this case, we will need to include additional variables in order to make the right decision.

**Remarks**

The above steps imply that in order to make the right decision in association reversal conditions, we will first need to know the scenario behind the dataset so that we can use the scenario to construct a causal graph and then decide whether or not the third variable Z satisfies the backdoor criterion. Therefore, it shows that the use of causal inference here is not "free-lunch", we will always require *additional information* (here the scenario behind the data) in order to improve our understanding of the data.

## 3.2.4 Understanding Simpson's paradox in Example 3 using causal graphical models

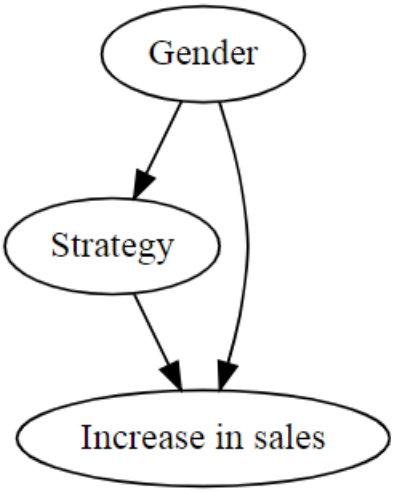
### Step 1: Construct a causal graph

Now suppose we are given some additional information on how the advertisement strategies were implemented:

- The marketing department decides to put advertisements only on sports TV channels (for the broadcast strategy) and online shopping websites (for the internet strategy).
- Most male customers spent much more time on sports TV channels than on online shopping websites, while most female customers spent much more time on online shopping websites than on sports TV channels [this tells us that the gender variable has a causal effect on the strategy variable]

- In general, female customers tend to spend more money on buying running shoes compared to male customers [this tells us that the gender variable has a causal effect on the increase in sales variable]
- Historical evidence has shown that different types of strategy had different impact on increase of sales [this tells us that the strategy variable has a causal effect on the increase in sales variable]

With the above additional information, the following causal graph can be constructed:




**Step 2: Find all backdoor paths using the causal graph**

Here, the only backdoor path is

$$Strategy \leftarrow \text{Gender} \rightarrow Increase\ in\ sales$$

**Step 3: Make the right decision using the backdoor criterion**

Now, the question becomes whether Gender satisfies the backdoor criterion relative to the variables (Strategy, Increase in sales).

In this path, Gender acts as a fork which means blocking Gender will block this backdoor path. In addition, because Gender is not a descendant of Strategy, we know that Gender does satisfy the backdoor criterion relative to (Strategy, Increase in sales). Therefore, in this case, the *segregated data* will give us the right decision. In other words, causal inference tells us that strategy #2 (broadcast advertisement) is the better strategy to use if we want to increase sales for both male and female customers and therefore it is the better strategy for the whole population.

**Step 4 Explain the paradox using expert knowledge or educated guess**

Here is one possible explanation of what is happening. In general, female customers tend to have a better chance of purchasing new running shoes than male customers (due to various reasons) and more likely to be affected by the internet strategy, while male customers are more likely to be affected by the broadcast strategy. Now, when we analyze the data using the whole population, the internet strategy seems to be better as it is overwhelmed by female customers which have a bigger impact on the increase in sales.

## 3.3 Birth-weight paradox and Berkson's paradox

In general, causal graphical models can be used to explain many paradoxes which are surprising to people if they do not have a good understanding of causal inference. This section discusses two other famous paradoxes: birth-weight paradox and Berkson's paradox.

**Example 4 Birth-weight paradox**

The birth-weight paradox is a famous paradox about the relationship between birth weight of children and their mortality rate based on their mother's smoking status. The paradox can be summarized as follows:

Suppose we want to study the relationship between birth weight and mortality rate of children. Past studies have shown that children of tobacco smoking mothers are more likely to give children with low birth weight, and children with low birth weight will have a much higher mortality rate than children with normal birth weight. However, when researchers analyzed the actual real-world data (that is, observational data), they found that low birth weight children of smoking mothers actually had lower mortality rate than low birth weight children of non-smoking mothers. This is very surprising to people as the observational data tells us that having a smoking mother is actually beneficial to a child's health! The birth-weight paradox is against our common intuition that smoking is harmful to one's health and if the mother really believes this conclusion then she will probably want to smoke more tobacco in order to have a healthy baby.

**Explain birth-weight paradox using causal graphical models**

**Step 1: Construct the causal graph**

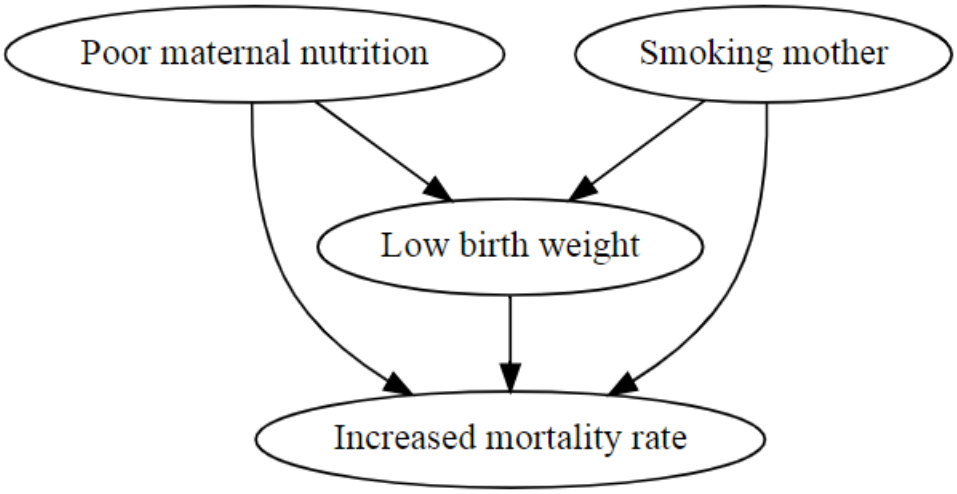
From the above story we can obtain the following information, that is:

- Smoking has a causal effect on birth weight
- Birth weight has a causal effect on mortality rate
- Smoking has a causal effect on mortality rate (as shown in many studies)

Suppose in addition, we also know the following fact:

- Another key factor that is related to both birth weight and mortality rate is maternal nutrition. Poor maternal nutrition is shown to cause low birth weight as well as higher mortality rate of children.
- We also believe that maternal nutrition is independent of tobacco smoking in general (this makes sense as there are smokers in both poor and wealthy groups)

Now, we can construct the following causal graph to represent the given information:




**Step 2: Make the right decision using the backdoor criterion**

First it should be noted that the paradox is really about the appearance of negative causal effect of *smoking mother* on *increased mortality rate of children* if conditioned on *low birth weight*. However, from the causal graph, we can see that there is actually *no backdoor path* that requires blockage in relative to variables (Smoking mother, Increased mortality rate). Therefore, when investigate the causal relationship between smoking and mortality rate, we should not condition

on any additional variable as this extraneous information will mislead us by introducing confounding bias.

**Step 3 Explain the paradox**

For the birth-weight paradox, one explanation can be that even though smoking is harmful and has a positive causal effect on mortality rate, poor maternal nutrition has an even *more severe* impact than smoking when comes to mortality rate of children. Here, when considering low birth weight children, filtering children whose mothers smoke will reduce the chance of seeing poor maternal nutrition (for example, if a mother smokes tobacco and has poor maternal nutrition, then her child is probably going to die so seeing a low birth weight child is even not possible). This causes us to see that when conditioning on low birth weight, smoking mothers tend to have lower mortality rate of their children.

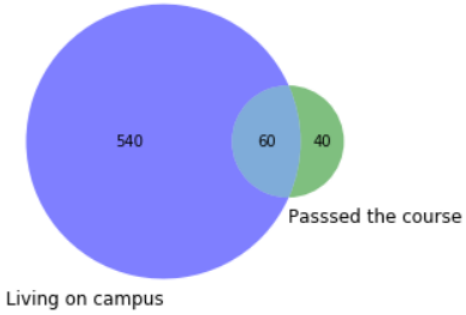
**Example 5 Berkson's paradox**

Berkson's paradox (also known as Berkson's bias or Berkson's fallacy) is a famous paradox that is caused by *sample selection bias* and is quite commonly seen in *self-selected sampling* problems. The paradox is illustrated through the following problem:

Suppose a college is offering a new course in causal inference and wants to find the relationship between a student's performance of the course and whether or not he or she lives on campus. In order to do so, the college did a survey in campus on Friday night and collected the following information:

- Among the surveyed students, 100 students passed the course
- Among the surveyed students, 600 students lived on campus
- Among the surveyed students,  60 students lived on campus AND passed the course
- In total, 640 students are being sampled

The Venn diagram gives a clear picture of the above information.




Using the obtained survey results, the college performed the following analysis:

$$Pass\ rate\ of\ sampled\ students = \frac{100}{640} \approx \mathbf{16}\%$$

$$Pass\ rate\ of\ sampled\ students\ who\ live\ on\ campus = \frac{60}{600} = \mathbf{10}\%$$

As a result, the college concludes that living on campus can cause a student's performance to be *worse* in the course.

But is this really true?

Now, suppose that there are 1,000 students in the population and we obtain the following information:

- Among the population, 100 students passed the course

- Among the population, 600 students lived on campus
- Among the population,  60 students lived on campus AND passed the course

Using the population information, we performed the following analysis:

$$Pass\ rate\ of\ students = \frac{100}{1000} = \mathbf{10}\%$$

$$Pass\ rate\ of\ students\ who\ live\ on\ campus = \frac{60}{600} = \mathbf{10}\%$$

As a result, we can conclude that there is no clear evidence that living on campus is related to a student's performance. So what causes the different conclusions?

**Explain the results from a sampling perspective**

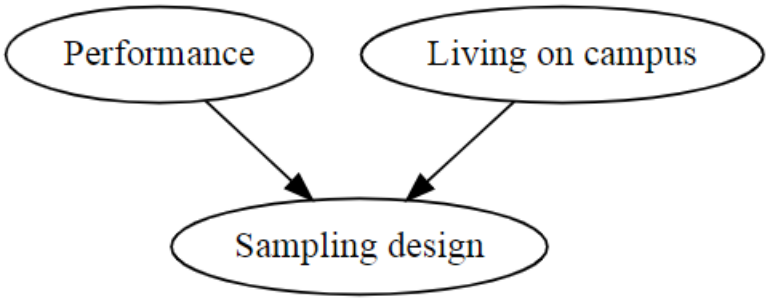
In this toy example, it is noticed that the college's sample only includes those students who either live on campus OR passed the course, and one explanation can be that a student who lives off campus AND failed the course all had no motivation to go campus on a Friday night and therefore the collected samples includes none of them. As a result, using the collected samples will distort the analysis by the so-called ***sample selection bias*** – that is, the sample is a non-random representation of the population and causing some units of the population to be less likely to be included than others. Making inference on such samples will cause characteristics of certain groups to be biased, in this case, the pass rate of students in the sampled data is inflated.

**Explain Berkson's paradox using a causal graph**

A careful look at the problem also introduces a counterintuitive result for some people, that is, given two independent events, only considering results where only at least one of these events occurs will make them to become correlated to each other. In this example, living on campus and pass the course are independent events, but when only considering one of them, they will seem to be correlated to each other. This is another example that correlation is not the same as causation, where using only statistical evidence alone can lead to wrong causal conclusions.

A causal graph can be constructed for this example using the following knowledge:

- There are three variables in the graph: Performance, Living on campus, and Sampling design
- Both Performance and Living on campus have causal effect on the Sampling design
- Performance and Living on campus are independent (therefore no connected edge between them)




As we can see, the example can be represented by a collider (or V-structure) graph. Since Sampling design acts as the collider, condition on it (in this case using such sampling design) will cause the path to be opened and make Performance to be dependent to Living on campus. This is true as living on campus indicates worse performance in the sample data that is obtained using the selected sampling design, but not in the population data.

# 4. Frontdoor adjustment and unobserved variables

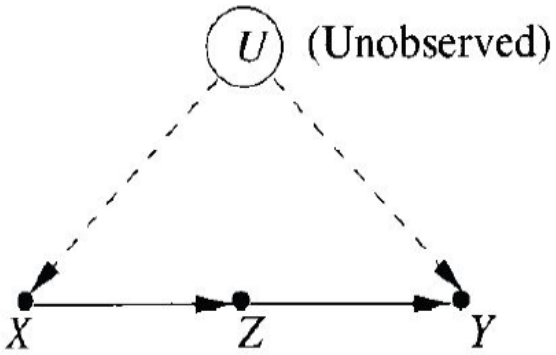
## 4.1 Unobserved variables and confounding effects

Simpsons' paradox shows us when there exists confounding variables which are common causes for both variables X and Y, then it is possible that the observed association between the variables can be either a direct causal relationship between them or due to the confounding variables through the backdoor paths. We see that if we attest that we have observed enough number of the confounding variables, then through a causal graph, we can make the right decision by using the backdoor criterion and adjustment formula. However, in many cases, obtaining information about all possible confounders is impossible due to various reasons. For example, randomized controlled trials are usually considered to be the most efficient method to assess treatment effects, however, in some cases, it is not legal or ethical to conduct randomized experiment. Consider the famous debate on the causal relationship between smoking and lung cancer, then we cannot simply encourage non-smokers to smoke or force smokers to quit smoking. In this case, we will need to apply another approach – the frontdoor adjustment to make causal inference. This section briefly summarizes the frontdoor adjustment approach and illustrates how it can be used to make the right decision through an example.

## 4.2 Frontdoor criterion and frontdoor adjustment formula

Consider the following causal graph with 4 variables X, Y, Z, and U. Here, X, Y, and Z are all observed while U is unobserved or latent. Since U is an unobserved confounder, we cannot use

the backdoor adjustment formula to make inference about the interventions. However, since we observed the mediator Z, we can use another approach for the problem.




From the graph, it can be seen that the joint distribution of (X, Y, Z, U) can be factorized as:

$$P(x, y, z, u) = P(u)P(x|u)P(z|x)P(y|z, u)$$

In addition, the graph also tells us the following conditional independence relationships:

1) $U \text{ and } Z \text{ are conditionally independent given } X$

$$<=> P(u|x) = P(u|z, x)$$

2) $X \text{ and } Y \text{ are conditionally independent given both } Z \text{ and } U$

$$<=> P(y|z, u) = P(y|x, z, u)$$

When we intervene on X, then the edge from U to X can be removed and we get:

$$P\big(y, z, u\big|do(x)\big) = P(u)P(z|x)P(y|z, u)$$

$$=> P\left(y\middle|do(x)\right) = \sum_{(z,u)} P(u)P(z|x)P(y|z,u)$$

$$= \sum_{z} P(z|x) \sum_{u} P(u)P(y|z,u)$$

Now, using the two conditional independence relationships, $\sum_u P(u)P(y|z,u)$ can be also written as:

$$\sum_{u} P(u)P(y|z,u)$$

$$= \sum_{u} P(u)P(y|x,z,u)\,, as\ X \perp \mathrm{Y} \mid (Z,U)$$

$$= \sum_{x}\sum_{u} [P(u|x)P(x)]P(y|x,z,u)$$

$$= \sum_{x}\sum_{u} P(u|z,x)P(x)P(y|x,z,u),\ as\ Z \perp \mathrm{U} \mid \mathrm{X}$$

$$= \sum_{x} P(x)P(y|x,z)\,, which\ now\ has\ no\ latent\ variable\ U\ in\ it$$

$$Therefore, \boldsymbol{P\left(y\middle|do(x)\right)}$$

$$= \sum_{z} P(z|x) \sum_{u} P(u)P(y|z,u)$$

$$= \sum_{\boldsymbol{z}} \boldsymbol{P(z|x)} \sum_{\boldsymbol{x'}} \boldsymbol{P(x')P(y|x',z)}\,, by\ changing\ the\ variable\ name\ x\ to\ x'$$

$$in\ the\ second\ summation\ sign$$

This is known as the ***frontdoor adjustment formula***.

**Remarks:**

- The above frontdoor adjustment formula gives us a way to estimate the causal effect of X on Y in scenario where an unobserved variable U is present under the case where the two conditional independence relationships $\boldsymbol{Z} \perp \boldsymbol{U} \mid \boldsymbol{X}$ and $\boldsymbol{X} \perp \mathbf{Y} \mid (\boldsymbol{Z}, \boldsymbol{U})$ hold true (i.e. when Z acts as a chain/mediator for X and Y and when the unobserved variable U only affects X and Y but not Z.
- The frontdoor adjustment formula can be thought as a two-step application of the backdoor adjustment formula through the chain $\boldsymbol{X} \rightarrow \boldsymbol{Z} \rightarrow \boldsymbol{Y}$
  - Step 1: Find the causal effect of X on Z. Note that since there is no backdoor path from X to Z, we know that the causal effect is the same as the correlation effect, in other words, $\boldsymbol{P}\big(\boldsymbol{z}\big|\boldsymbol{do}(\boldsymbol{x})\big) = \boldsymbol{P}(\boldsymbol{z}|\boldsymbol{x})$, which is exactly the first piece in the frontdoor adjustment formula.
  - Step 2: Find the causal effect of Z on Y. Note that there is one backdoor path $\boldsymbol{Z} \leftarrow \boldsymbol{X} \leftarrow \boldsymbol{U} \rightarrow \boldsymbol{Y}$ and we know that X here satisfies the backdoor criterion (as block X will block this backdoor path and X is not a descendant of Z). Then the backdoor adjustmen formula tells that the causal effect of Z on Y can be computed as:

$$\boldsymbol{P}\big(\boldsymbol{y}\big|\boldsymbol{do}(\boldsymbol{z})\big) = \sum_{\boldsymbol{x'}} \boldsymbol{P}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x'})\boldsymbol{P}(\boldsymbol{x'}) \text{ , by replacing the variable name x with x}',$$

which is exactly the second piece in the frontdoor adjustment formula.

**Definition 9 (Frontdoor criterion)**

Let $G(V, E)$ be a causal Bayesian network with observational data on the set of variables V. A set of variables $Z \subseteq V$ satisfies the frontdoor criterion relative to variables $(X, Y)$ if Z satisfies the following conditions:

1. Z intercepts all directed paths from X to Y
2. There is no backdoor path from X to Z
3. All backdoor paths from Z to Y are blocked by X

**Remarks:**

- In condition 1, the set of variables Z acts as a *mediator* for all directed paths from X to Y
- Condition 2 implies that $\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{do}(\boldsymbol{x})) = \boldsymbol{P}(\boldsymbol{z}|\boldsymbol{x})$ [correlation is the same as causation]
- Since it is also true that X can never be a descendant of Z (as Z is a mediator for paths from X to Y), condition 3 implies that X satisfies the backdoor criterion relative to variables (Z,Y)
- For example, in the above example, variable Z satisfies the frontdoor criterion while the empty set does not

**Theorem 6 (Frontdoor adjustment formula)**

Given a set of variables Z which satisfies the frontdoor criterion relative to variables $(X, Y)$ and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the following formula:

$$\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{do}(\boldsymbol{z})) = \sum_{\boldsymbol{z}} \boldsymbol{P}(\boldsymbol{z}|\boldsymbol{x}) \sum_{\boldsymbol{x'}} \boldsymbol{P}(\boldsymbol{x'})\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x'}, \boldsymbol{z})$$

**Example 6 Unmeasured confounders and frontdoor adjustment formula**

Suppose we want to study the causal relationship between lack of sleep (X) and risk of sudden death (Y). Past study has shown that lack of sleep can cause higher risk of heart attack (Z), and higher risk of heart attack will cause higher risk of sudden death. Suppose the following observational data is obtained:

- X = 0 means enough sleep and X = 1 means lack of sleep
- Y = 0 means never experienced any sign of sudden death and Y = 1 means experienced some signs of sudden death
- Z = 0 means never experienced heart attack and Z = 1 means experienced heart attack

| **Group** | **(X,Z)** | **P(x,z)** | $\mathbf{P(Y=1|x,z)}$ |
|---|---|---|---|
| Enough sleep<br>No heart attack | (0,0) | 0.45 | 0.1 |
| Lack of sleep<br>No heart attack | (1,0) | 0.05 | 0.1 |
| Enough sleep<br>Experience heart attack | (0,1) | 0.05 | 0.8 |
| Lack of sleep<br>Experience heart attack | (1,1) | 0.45 | 0.1 |

**Analysis 1: assume no unmeasured confounder exists**

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=0)=\frac{P(Y=1,X=0)}{P(X=0)}=\frac{P(Y=1,X=0,Z=0)+P(Y=1,X=0,Z=1)}{P(X=0)}$$

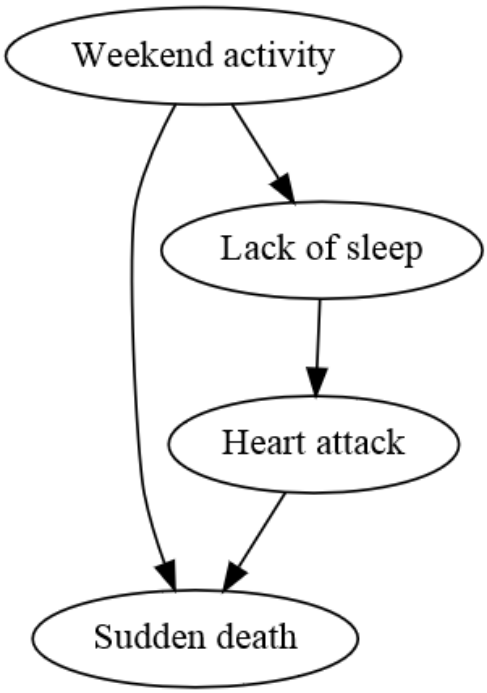
$$=\frac{0.45*0.1+0.05*0.8}{0.5}=\mathbf{0.17}$$

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=1)=\frac{P(Y=1,X=1)}{P(X=1)}=\frac{P(Y=1,X=1,Z=0)+P(Y=1,X=1,Z=1)}{P(X=0)}$$

$$=\frac{0.05*0.1+0.45*0.1}{0.5}=\mathbf{0.10}$$

The result actually shows that getting enough sleep will cause a higher risk of sudden death (which is against our common intuition).

Now, suppose we found that what people like to do during the weekends is also a key contributor to both the risk of sudden death and lack of sleep but does not have a direct impact on risk of heart disease, denote this variable Weekend activity U (U=0 means study during weekends and U=1 means go to nightclub during weekends). However, the collected samples do not have any information on this variable, in other words, the variable Weekend activity is an unobserved confounder. The following causal graph can be used to represent the above scenario:




Note that in this case, we cannot use the backdoor adjustment formula to evaluate the treatment effect as the variable Weekend activity is unobserved. However, we know that variable Z (the risk of heart disease) satisfies the frontdoor criterion, therefore, we can apply the frontdoor adjustment formula to compute the treatment effect of enough sleep and lack of sleep.

**Analysis 2: consider the unmeasured confounder**

$\mathbf{P(Y = 1|do(X = 0))}$

$= P(Z = 0|X = 0) * [P(Y = 1|Z = 0, X' = 0) * P(X' = 0) + P(Y = 1|Z = 0, X' = 1) * P(X' = 1)] +$

$P(Z = 1|X = 0) * [P(Y = 1|Z = 1, X' = 0) * P(X' = 0) + P(Y = 1|Z = 1, X' = 1) * P(X' = 1)]$

$$= \frac{0.45}{(0.45+0.05)} * [0.1 * 0.5 + 0.1 * 0.5] + \frac{0.05}{(0.45+0.05)} * [0.8 * 0.5 + 0.1 * 0.5] = 0.09 + 0.045$$

$$= \mathbf{0.135}$$

$$\mathbf{P(Y = 1|do(X = 1))}$$

$$= P(Z = 0|X = 1) * [P(Y = 1|Z = 0, X' = 0) * P(X' = 0) + P(Y = 1|Z = 0, X' = 1) * P(X' = 1)] +$$

$$P(Z = 1|X = 1) * [P(Y = 1|Z = 1, X' = 0) * P(X' = 0) + P(Y = 1|Z = 1, X' = 1) * P(X' = 1)]$$

$$= \frac{0.05}{(0.45+0.05)} * [0.1 * 0.5 + 0.1 * 0.5] + \frac{0.45}{(0.45+0.05)} * [0.8 * 0.5 + 0.1 * 0.5] = 0.01 + 0.405$$

$$= \mathbf{0.415}$$

By considering the unmeasured confounder, we find that getting enough sleep will cause a much lower risk of sudden death which is exactly the opposite conclusion compared to the case where the unmeasured confounder is ignored.

**Remarks**

The above example shows how the frontdoor adjustment approach can be used to compute the treatment effects even when we do not have any observational data on the unmeasured confounder. Again, similar to the backdoor adjustment case, a prior knowledge about the scenario is needed in order to construct the causal graph, and coupled with the observational data, we can then compute the treatment effects and make the proper causal conclusion. Finally,

through this fake data, we see how confounding bias can be dangerous in making causal conclusions, lack of knowledge on the scenario can mislead us into completely wrong decisions.

## 5. Conclusion

In summary, this report illustrates how causal graphical models can be used with observational data to make causal conclusions using the backdoor and frontdoor adjustment methods. Through various simulated data examples, we see that statistical paradoxes exist because association does not imply causation, and making causal conclusions solely based on statistical results can be counterintuitive and misleading. Finally, it should be noted that causal inference is not magic, in order to make the correct causal conclusions, prior expert knowledge on the underlying data-generating process is needed so that a causal graph can be constructed to represent the independence structure of the data.

# Appendix (Python code)

### 1) Construct the three basic structures of a Bayesian network

```
from causalgraphicalmodels import CausalGraphicalModel
import causalgraphicalmodels
from causalgraphicalmodels.examples import fork, chain, collider
chain.draw()
fork.draw()
collider.draw()
```

### 2) Illustrate the concept of d-separation

```
#Create and draw the Bayes net
path = CausalGraphicalModel(
    nodes = ["x1", "x2", "x3", "x4", "x5","x6","x7"],
    edges = [("x1", "x6"), ("x1", "x2"), ("x1", "x3"), ("x2", "x4"), ("x3", "x5"), ("x4", "x6"), ("x5", "x6"),("x6", "x7")]
)
path.draw()


#Confirm whether the following conditional independence relationships are true or not
path.is_d_separated("x1", "x4", {"x2"})
path.is_d_separated("x1", "x4", {"x2","x6"})
path.is_d_separated("x1", "x4", {"x2","x7"})
path.is_d_separated("x1", "x4", {})
```

### 3) Illustrate the concept of backdoor criterion

```
#Create and draw the Bayes net
path = CausalGraphicalModel(
    nodes = ["x1", "x2", "x3", "x4", "x5","x6","x7","x8","x9","x10"],
    edges = [("x2", "x4"), ("x2", "x5"), ("x3", "x5"), ("x4", "x6"), ("x5", "x7"), ("x6", "x8"),("x7", "x9"),
             ("x8", "x10"),("x9", "x10"),("x3", "x5"),("x1", "x10"),("x3", "x10"),("x4", "x9"),("x5", "x8")]
)
path.draw()

#Find all backdoor paths from x5 to x9
path.get_all_backdoor_paths("x5", "x9")

#Get all sets of variables that satisfy the backdoor criterion
path.get_all_backdoor_adjustment_sets("x5", "x9")
```

### 4) Simpson's paradox

```
#Frist simulate the dataset
random.seed(1234) #set the random seed
#Simulate Z
Z = random.binomial(n=1, p=0.5, size=2000)
#Simulate X based on Z
X = random.binomial(n=1, p=9/10-4/5*Z, size=2000) + 1
#Simulate Y based on X and Z
mu, sigma = X+2*Z, 2
Y = random.normal(mu, sigma, size=2000)

#Converts X, Y, Z into a dataframe
from pandas import DataFrame
df1 = DataFrame({'Advertising strategy':X, 'Increase in sales (%)': Y})
```

```
df2 = DataFrame({'Gender':Z, 'Advertising strategy':X, 'Increase in sales (%)': Y})
###################
#1. Perform analysis using the aggregated data (df1)
sns.boxplot(x = 'Advertising strategy', y = 'Increase in sales (%)', data=df1)
plt.title("Increase in sales (%) vs Advertising strategy")
plt.show()

sns.violinplot(x = 'Advertising strategy', y = 'Increase in sales (%)', data=df1)
plt.title("Increase in sales (%) vs Advertising strategy")
plt.show()

df1.groupby('Advertising strategy').mean()
###################
#2. Perform analysis using the segragated data (df2)
sns.boxplot(x='Advertising strategy', y='Increase in sales (%)', data=df2[df2['Gender'] == 0])
plt.title("Increase in sales (%) vs Advertising strategy for Males")
plt.show()
sns.boxplot(x='Advertising strategy', y='Increase in sales (%)', data=df2[df2['Gender'] == 1])
plt.title("Increase in sales (%) vs Advertising strategy for Females")
plt.show()

sns.violinplot(x='Advertising strategy', y='Increase in sales (%)', data=df2[df2['Gender'] == 0])
plt.title("Increase in sales (%) vs Advertising strategy for Males")
plt.show()
sns.violinplot(x='Advertising strategy', y='Increase in sales (%)', data=df2[df2['Gender'] == 1])
plt.title("Increase in sales (%) vs Advertising strategy for Females")
plt.show()

df2.groupby(['Gender', 'Advertising strategy']).mean()
###################
#3. Decide the right decision using a causal graph
```

```
#Create and draw the causal graph
path = CausalGraphicalModel(
    nodes = ["Gender", "Strategy", "Increase in sales"],
    edges = [("Gender", "Strategy"), ("Gender", "Increase in sales"), ("Strategy", "Increase in sales")]
)
path.draw()

#Obtain the backdoor path
path.get_all_backdoor_paths("Strategy", "Increase in sales")

#Whether Gender satisfies the backdoor criterion
path.is_valid_backdoor_adjustment_set("Strategy", "Increase in sales", {"Gender"}) #TRUE
```

**5) Birth-weight paradox**

```
#Create and draw the causal graph
path = CausalGraphicalModel(
    nodes = ["Smoking mother", "Low birth weight", "Increased mortality rate","Poor maternal nutrition"],
    edges = [("Smoking mother", "Increased mortality rate"), ("Smoking mother", "Low birth weight"),
             ("Low birth weight", "Increased mortality rate"),("Poor maternal nutrition","Low birth weight"),
             ("Poor maternal nutrition","Increased mortality rate")]
)
path.draw()

#Obtain the backdoor path
path.get_all_backdoor_paths("Smoking mother", "Increased mortality rate")
```

### 6) Berkson's selection bias

```
#Illustrate the data using a Venn diagram
v = vplt.venn2(subsets = {'10':540,'01':40,'11':60},set_labels=('Living on campus','Passsed the course'),
    set_colors=('b','g'),alpha=0.5)
plt.show()


#Create and draw the causal graph
path = CausalGraphicalModel(
    nodes = ["Performance", "Living on campus", "Sampling design"],
    edges = [("Performance", "Sampling design"), ("Living on campus", "Sampling design")]
)
path.draw()


#obtain the backdoor path
path.get_all_backdoor_paths("Performance", "Living on campus")
```

### 7) Unmeasured confounder and frontdoor adjustment

```
#Create and draw the causal graph
path = CausalGraphicalModel(
    nodes=["Lack of sleep", "Heart attack", "Sudden death","Weekend activity"],
    edges=[("Lack of sleep", "Heart attack"), ("Heart attack", "Sudden death"),
                ("Weekend activity", "Lack of sleep"),("Weekend activity", "Sudden death")],
)
path.draw()
```

# References

Pearl, J. (2014). Understanding Simpson's Paradox. The American Statistician. Retrieved November 20, 2018.

Aussem, A. Data Mining & Machine Learning (DM2L) Group. Université Claude Bernard Lyon 1. Causal Inference & Paradoxes. LIRIS UMR 5205 CNRS. Retrieved November 20, 2018.

Mohan, K., Pearl, J. (2014). Graphical Models for Causal Inference. University of California, Los Angeles. Retrieved November 20, 2018.

PEARL, J. (2009). CAUSALITY - 3.3 Controlling confounding bias (2nd ed.).

Bayesian network. (2018, October 11). Retrieved November 20, 2018, from https://en.wikipedia.org/wiki/Bayesian_network