Causality

"According to studies..."

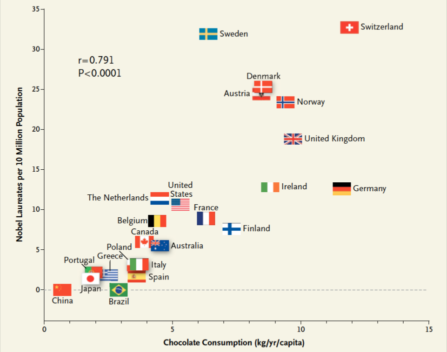

Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- What would be the right question in this setting?

- What would be the right question in this setting?
- Numerous studies concern themselves *only* with correlation

- What would be the right question in this setting?
- Numerous studies concern themselves *only* with correlation
- This can be very misleading

- What would be the right question in this setting?
- Numerous studies concern themselves *only* with correlation
- This can be very misleading
- `http://tylervigen.com/spurious-correlations`

Simpson's Paradox

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Figure: Success rates of two treatments for kidney stones

- Treatment B seems to perform better overall (83%)

Simpson's Paradox

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Figure: Success rates of two treatments for kidney stones

- Treatment B seems to perform better overall (83%)
- **But** treatment A performs better in both settings

Posing correct questions

- Correlation vs. Causality

Posing correct questions

- Correlation vs. Causality
- Missing background knowledge can lead to false conclusions

Posing correct questions

- Correlation vs. Causality
- Missing background knowledge can lead to false conclusions
- Correlation does **not** imply causality

Posing correct questions

- Correlation vs. Causality
- Missing background knowledge can lead to false conclusions
- Correlation does **not** imply causality
- Mostly we're interested if A is having a direct effect on (causing) B

Posing correct questions

- Correlation vs. Causality
- Missing background knowledge can lead to false conclusions
- Correlation does **not** imply causality
- Mostly we're interested if A is having a direct effect on (causing) B
- What are possible causal explanations if A is correlated to B?

Posing correct questions

- Correlation vs. Causality
- Missing background knowledge can lead to false conclusions
- Correlation does **not** imply causality
- Mostly we're interested if A is having a direct effect on (causing) B
- What are possible causal explanations if A is correlated to B?
  - i) A causes B, ii) B causes A or iii) hidden actor Z causes A and B

Posing correct questions

- Correlation vs. Causality
- Missing background knowledge can lead to false conclusions
- Correlation does **not** imply causality
- Mostly we're interested if A is having a direct effect on (causing) B
- What are possible causal explanations if A is correlated to B?
  - i) A causes B, ii) B causes A or iii) hidden actor Z causes A and B
  - Reichenbachs common cause principle is provable

Hidden Cause/Actor

- In 1999 research established a significant correlation between the presence of a nightlight in a child's bedroom and myopia (shortsightedness).

Hidden Cause/Actor

- In 1999 research established a significant correlation between the presence of a nightlight in a child's bedroom and myopia (shortsightedness).
- In 2000 follow-up research found out that parents with myopia are more likely to put a nightlight in their child's bedroom. Their children also are more inclined to develop myopia for genetical reasons.

Motivation

"Correlation does not imply causation" $\Leftrightarrow$ $\mathbb{P}^{\mathbf{X}} \neq \tilde{\mathbb{P}}^{\mathbf{X}}$

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.
- i is a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$, i.e. j is a child of i

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.
- i is a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$, i.e. j is a child of i
- an edge is **undirected** if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$, otherwise it is **directed**

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.
- i is a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$, i.e. j is a child of i
- an edge is **undirected** if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$, otherwise it is **directed**
- 3 nodes form an **immorality (or v-structure)** if one is the child of the two others that themselves are not adjacent

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.
- i is a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$, i.e. j is a child of i
- an edge is **undirected** if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$, otherwise it is **directed**
- 3 nodes form an **immorality (or v-structure)** if one is the child of the two others that themselves are not adjacent
- a **(directed) path** is a sequence of distinct $i_1, \dots, i_n \in V$ with a (directed) edge between i_k and i_{k+1} for all $k = 1, \dots, n-1$

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.
- i is a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$, i.e. j is a child of i
- an edge is **undirected** if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$, otherwise it is **directed**
- 3 nodes form an **immorality (or v-structure)** if one is the child of the two others that themselves are not adjacent
- a **(directed) path** is a sequence of distinct $i_1, \dots, i_n \in V$ with a (directed) edge between i_k and i_{k+1} for all $k = 1, \dots, n-1$
- all j with a directed path from i to j are called **descendants** of i, the set of all descendants of i is denoted by $DE_i^{\mathcal{G}}$

Graphs

- A **graph** $\mathcal{G} = (V, \mathcal{E})$ consists of nodes V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E} \quad \forall v \in V$.
- i is a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$, i.e. j is a child of i
- an edge is **undirected** if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$, otherwise it is **directed**
- 3 nodes form an **immorality (or v-structure)** if one is the child of the two others that themselves are not adjacent
- a **(directed) path** is a sequence of distinct $i_1, \dots, i_n \in V$ with a (directed) edge between i_k and i_{k+1} for all $k = 1, \dots, n-1$
- all j with a directed path from i to j are called **descendants** of i, the set of all descendants of i is denoted by $DE_i^{\mathcal{G}}$
- we identify the nodes $j \in V$ with the random variables $X_j \in \mathbf{X}$

DAGs

- a **directed acyclic graph (DAG)** is $\mathcal{G}$ in which there exists no (i, j) with directed paths from i to j and from j to i, and all the edges are directed

DAGs

- a **directed acyclic graph (DAG)** is $\mathcal{G}$ in which there exists no (i, j) with directed paths from i to j and from j to i, and all the edges are directed
- in a DAG, the disjoint $A, B \subset V$ are *d*-**separated** by a also disjoint $S \subset V$ if every path between nodes in A and B is blocked by S, i.e. for every path i_1 to i_n:
  - $i_k \in S$ and $i_{k-1} \rightarrow i_k \rightarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \rightarrow i_{k+1}$
  - $i_{k-1} \rightarrow i_k \leftarrow i_{k+1}$ and neither i_k nor any of its descendants is in S

Topological Ordering

Proposition:

For each DAG exists a **topological ordering** $\pi \in S_p$, that is a bijective mapping

$$\pi : \{1, ..., p\} \to \{1, ..., p\}$$

that satisfies

$$\pi(i) < \pi(j) \quad \text{if} \quad j \in DE_i^{\mathcal{G}}$$

Structural Equation Model

Definition:

A **structural equation model (SEM)** is $\mathcal{S} := (\mathbf{S}, \mathbb{P}^N)$, where $\mathbf{S} = (S_1, \ldots, S_p)$ are equations

$$S_j : \quad X_j = f_j(PA_j, N_j), \qquad j = 1, \ldots, p$$

Interventions

Having established the SEM structure, we now can construct new distributions by changing (intervening upon) structural equations.

Definition (Intervention Distribution)

Consider the distribution SEM $(\mathcal{S}, \mathbb{P}^{\mathbf{N}}) \rightsquigarrow \mathbb{P}^{\mathbf{X}}$. We now can replace one or multiple equations and obtain a new SEM $\widetilde{\mathcal{S}}$. The new distribution $\mathbb{P}^{\mathbf{N}}_{\tilde{\mathcal{S}}}$ is called the **intervention distribution** and the variables whose structural equations have been changed have been intervened on. We introduce the *do* operator:

$$\mathbb{P}^{\mathbf{X}}_{\widetilde{\mathcal{S}}} =: \mathbb{P}^{\mathbf{X}|do(X_j=\tilde{f}(\tilde{\mathbf{PA}}_j,\tilde{N}_j))}_{\mathcal{S}}$$

Example for an Intervention (Kidney Stones)

- New and old N's need to be independent

- New and old N's need to be independent
- Two special cases

- New and old N's need to be independent
- Two special cases
  - The new equation can either keep the same parents but change their influence or restructure the noise component (called imperfect)

- New and old N's need to be independent
- Two special cases
  - The new equation can either keep the same parents but change their influence or restructure the noise component (called imperfect)
  - The new equation is of the type $do(X_j = a)$ (called perfect)

- New and old N's need to be independent
- Two special cases
  - The new equation can either keep the same parents but change their influence or restructure the noise component (called imperfect)
  - The new equation is of the type $do(X_j = a)$ (called perfect)
- Example: Suppose $\mathcal{S}$ is

$$X = N_X$$

$$Y = 4 \cdot X + N_Y$$

with $N_X, N_Y \sim \mathcal{N}(0, 1)$ Compare the intervention distribution of Y for $do(X = 2)$ and $do(X = 3)$ with $\mathbb{P}_{\mathcal{S}}^{Y}$? Now reverse the roles of X and Y. What happens?

Causal Effect

Definition (total causal effect)

Given a SEM $\mathcal{S}$

$$X \text{ has a causal effect on } Y \Leftrightarrow X \not\perp\!\!\!\perp Y \text{ in } \mathbb{P}_{\mathcal{S}}^{\mathbf{X}|do(X=\tilde{N}_X)}$$

TFAE:

- There is a causal effect X to Y

Causal Effect

Definition (total causal effect)

Given a SEM $\mathcal{S}$

$$X \text{ has a causal effect on } Y \Leftrightarrow X \not\perp\!\!\!\perp Y \text{ in } \mathbb{P}_{\mathcal{S}}^{\mathbf{X}|do(X=\tilde{N}_X)}$$

TFAE:

- There is a causal effect X to Y
- There are a, b s.t. $\mathbb{P}_{\mathcal{S}}^{Y|do(X=a)} \neq \mathbb{P}_{\mathcal{S}}^{Y|do(X=b)}$
- There is an a s.t. $\mathbb{P}_{\mathcal{S}}^{Y|do(X=a)} \neq \mathbb{P}_{\mathcal{S}}^{Y}$
- $X \not\perp\!\!\!\perp Y$ in $\mathbb{P}_{\mathcal{S}}^{X,Y|do(X=\tilde{N}_X)}$ for any $\tilde{N}_X$ whose dist. has full support

Remark:

- If there is no directed path from X to Y, then there is no causal effect
- Sometimes there is a directed path, but no causal effect.

Definition (Markov Property & Theorem)

Given a DAG G and a joint distribution $\mathbb{P}^{\mathbf{X}}$, this distribution is said to satisfy

- the **global Markov property** with respect to G if
  A,B d-sep. by C $\Rightarrow$ A $\perp\!\!\!\perp$ B | C $\forall$ disjoint sets A,B,C

Definition (Markov Property & Theorem)

Given a DAG G and a joint distribution $\mathbb{P}^{\mathbf{X}}$, this distribution is said to satisfy

- the **global Markov property** with respect to G if
  A,B d-sep. by C $\Rightarrow$ A $\perp\!\!\!\perp$ B | C $\forall$ disjoint sets A,B,C
- the **local Markov property** with respect to G if each variable is independent of its non-descendants given its parents

Definition (Markov Property & Theorem)

Given a DAG G and a joint distribution $\mathbb{P}^{\mathbf{X}}$, this distribution is said to satisfy

- the **global Markov property** with respect to G if
  A,B d-sep. by C $\Rightarrow$ A $\perp\!\!\!\perp$ B | C $\forall$ disjoint sets A,B,C
- the **local Markov property** with respect to G if each variable is independent of its non-descendants given its parents
- the **Markov factorization property** with respect to G if

$$p(\mathbf{x}) = p(x_1, ..., x_p) = \prod_{j=1}^{p} p(x_j | x_{\mathbf{PA}_j^G})$$

Definition (Markov Property & Theorem)

Given a DAG G and a joint distribution $\mathbb{P}^{\mathbf{X}}$, this distribution is said to satisfy

- the **global Markov property** with respect to G if
  A,B d-sep. by C $\Rightarrow$ A $\perp\!\!\!\perp$ B | C $\forall$ disjoint sets A,B,C
- the **local Markov property** with respect to G if each variable is independent of its non-descendants given its parents
- the **Markov factorization property** with respect to G if

$$p(\mathbf{x}) = p(x_1, ..., x_p) = \prod_{j=1}^{p} p(x_j | x_{\mathbf{PA}_j^G})$$

- **IF** $\mathbb{P}^{\mathbf{X}}$ has a densitiy p (w.r.t. a product measure), then all Markov properties above are **equivalent**!

Reichenbach's common cause principle can be proven using the previous Definitions and Theorem.

Proposition:

Assume that any pair of variables X and Y can be embedded into a larger system in the following sense: there exists a correct SEM over the collection **X** of random variables that contains X and Y with graph G. Then the Reichenbach's common cause principle follows from the Markov property in the following sense: If X and Y are dependent, then there is

- either a directed path from X to Y
- or from Y to X
- or there's a node T with a directed path from T to X and from T to Y.

Example:

Let the decision to study in Zurich ($Z = 1$) be determined only by whether one likes nature ($N = 1$) and whether one thinks ETH is a solid university ($U = 1$). How could the SEM look?

Example:

Let the decision to study in Zurich ($Z = 1$) be determined only by whether one likes nature ($N = 1$) and whether one thinks ETH is a solid university ($U = 1$). How could the SEM look?

- $N = N_N$

Example:

Let the decision to study in Zurich ($Z = 1$) be determined only by whether one likes nature ($N = 1$) and whether one thinks ETH is a solid university ($U = 1$). How could the SEM look?

- $N = N_N$
- $U = N_U$

Example:

Let the decision to study in Zurich ($Z = 1$) be determined only by whether one likes nature ($N = 1$) and whether one thinks ETH is a solid university ($U = 1$). How could the SEM look?

- $N = N_N$
- $U = N_U$
- $Z = (N \vee U) \oplus N_Z$

Example:

Let the decision to study in Zurich ($Z = 1$) be determined only by whether one likes nature ($N = 1$) and whether one thinks ETH is a solid university ($U = 1$). How could the SEM look?

- $N = N_N$
- $U = N_U$
- $Z = (N \vee U) \oplus N_Z$
- choose $N_N, N_U \sim^{iid}$ Ber(0.5) and $N_Z \sim^{iid}$ Ber(0.1)

From the SEM we can see that N and U are assumed to be independent. If you ask engineering students in Zurich (you condition on $Z = 1$, the answers to whether they like nature or think that ETH is a good university become anti-correlated: if someone is not a fan of nature, he probably likes ETH and vice versa. (Else he'd probably not have studied at ETH due to Ber(0.1)). So we have

$$N \not\perp\!\!\!\perp U \,|\, (Z = 1).$$

Truncated Factorization

Consider SEM $\mathcal{S}$ with structural equations

$$X_j = f_j(X_{pa(j)}, N_j)$$

and density $p_{\mathcal{S}}$. We have

$$p_{\mathcal{S}}(x_1, \ldots, x_p) = \prod_{j=1}^{p} p_{\mathcal{S}}(x_j | x_{pa(j)})$$

Truncated Factorization

Construct $\tilde{\mathcal{S}}$ from $\mathcal{S}$ by $do(X_k = \tilde{N}_k)$

$$p_{\mathcal{S},do(X_k=\tilde{N}_k)}(x_1,\dots,x_p) = \prod_{j=1}^{p} p_{\mathcal{S},do(X_k=\tilde{N}_k)}(x_j|x_{pa(j)}) = \prod_{j\neq k} p_{\mathcal{S}}(x_j|x_{pa(j)})\tilde{p}(x_k)$$

Special Case:

$$\mathrm{p}_{\mathcal{S},do(X_k=a)}(x_1,\dots,x_p) = \begin{cases} \prod_{j\neq k} p_{\mathcal{S}}(x_j|x_{pa(j)}) & \text{if } x_k = a \\ 0 & \text{otherwise} \end{cases}$$

References

Jonas Peters (2015). Causality, lecture notes