UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Introducción a la Inferencia Causal

Laura Cacheiro

Directora: Dra Mariela Sued

Fecha de Presentación: 14 de Julio de 2011

A la memoria de Martha y Ricardo

AGRADECIMIENTOS

Gracias[1]a Marie, por "hacerme ver", por su paciencia, por sus sonrisas, por demostrar que se puede ser inteligente, sencilla, sensible, y comprometida, por haber sido un punto de inflexión en mi carrera.
Gracias a Migue, por su amor, por no permitirme caer, por su entusiasmo contagioso, y todas sus envidiables cualidades sin las cuales no hubiera podido...
Gracias a mis solcitos Maria Azul y Manuela.

Gracias a Pati por las sugerencias y a Andrea Bergesio por las correcciones.

Gracias a todos los profesores, docentes auxiliares, que colaboraron en mi formación. En particular, gracias a los que mas influyeron: a Norberto Fava por sus clases perfectas y su trato siempre afectuoso; a Ursula Molter por el fervor que manifiesta cuando explica y por hacerme el aguante, en los primeros años de mi carrera ; a Eduardo Dubuc por maravillarme con sus enfoques matemáticos y a Gabo Mindlin por abrir mi cabeza, que no es poco decir, hacia los sistemas dinámicos y el caos.
A Mariano De Leo, Juan Pablo Borgna y Daniela Rodriguez, por soportar mis imbancables "no me sale", por atender mis consultas fuera de horario, entendiendo, lo difícil que es estudiar y trabajar.

A Pablo Solerno, por todo,todo,todo,todo,todo,todo,todo,todo,todo,todo,todo.
A Florencia Sember, por su apoyo, su cariño, sus carpetas y sus libros.
A Mariana Mazzón por todos los años de amistad y vivencias compartidos.

A Mercedes Fernandez Sau[2], Analia Lagorio, Enrique Di Rico, Andrés Muñoz, Laura Noni, Sebas Sosa, Ale Weil, Anita Ferrari, Tato Alvarez, Jorge Endelli, por compartir el estudio, los apuntes, las buenas ideas, el carnet de biblioteca, las malas ideas, las risas, los cafes, los bajones, las lapiceras, los nervios. A Guillermo Herrmann, por sus sugerencias acerca de este trabajo y nuestras charlas.

A mis amigas: Ceci, por leer "esto" a pesar de no ser matemática, Lore y Amy, por estar siempre.

A mis compañeros de Sinergia, por las horas ganadas pensando y tratando de hacer una facultad un poquito mejor.

A los 28 dias anuales por examen del GCBA.

[1] Esta palabra no transmite lo que me gustaría, no es el mismo gracias que digo veinte veces al día; estos GRACIAS se escriben de la misma manera pero significan algo muchísimo mas profundo, quizás una primera aproximación rockera podría ser GRACIAS TOTALES.

[2] mi socia :)

Índice general

Capítulo 1

Introducción

Este trabajo procura introducir y desarrollar algunos conceptos básicos en la inferencia causal.

Uno de los objetivos de la inferencia causal es identificar parámetros asociados a distribuciones de variables que no son observadas en todos los individuos de la población. Este hecho requiere de la aplicación de herramientas desarrolladas en el contexto de datos faltantes.

La causalidad es un área de interés en diferentes disciplinas, se trata de un concepto filosófico. Su complejidad se debe a que procura establecer afirmaciones sobre lo que no (necesariamente) sucedió, o lo que hubiera pasado si alguna circunstancia hubiera sido diferente. Un concepto causal es una relación que no puede ser definida a partir solamente de la distribución conjunta de las variables observadas. Luego las relaciones causales requieren introducir variables que ayuden a conceptualizar el problema de interés.

Si bien puede resultar dificultoso dar definiciones claras en términos coloquiales, la matemática brinda un rico soporte para abordar estas preguntas de manera sistemática.

El objetivo del análisis estadístico estándar es inferir parámetros de una distribución, a partir de muestras de la misma; estudiar asociaciones estadísticas, para lo cual utiliza típicamente probabilidad y técnicas de estimación, pero no hace una interpretación causal de los resultados. Sin embargo, muchas de las ciencias que utilizan la estadística vieron la necesidad de responderse preguntas de índole causal, como :¿Cuál es la eficacia de una droga en una población determinada ?¿Qué porcentaje de crímenes del pasado podrían haberse evitado con una política determinada? ¿Hace la obtención de un título universitario aumentar los ingresos de un individuo en el mercado laboral? Este es el tipo de preguntas causales a las que procura dar respuesta la inferencia causal.

La causalidad trasciende la matemática, resultando de suma importancia en diversas áreas del conocimiento. En particular, J.Robins en epidemiología, Heckman y C. Manski en economía, S.Morgan y R.Berk en sociología son algunos autores que han brindado un desarrollo propio, enriqueciendo y retroalimentando la causalidad. Cabe tambien mencionar a D.Lewis y W. Salmon dentro de la filosofía y J Pearl dentro de la computación científica. En parte a esto se debe la variablidad notacional con la que tendremos que lidiar en el presente trabajo.

Esta tesis está organizada de la siguiente manera.

Empezaremos en el Capítulo 2 por presentar un posible abordaje para estudiar efectos causales, incluyendo la noción de respuestas potenciales (o contrafactuales) y las definiciones matemáticas formales necesarias para lograr dar solución a lo que llamaremos problema de "identificación" de los parámetros causales. Se definen conceptos como aleatorización y aleatorización condicional y una interpretación "artesanal" con matemática básica de estas ideas.

En el Capitulo 3 mostramos como las funciones de probabilidad pueden ser asociadas a grafos de forma tal que condiciones de independencia o independencia condicional pueden ser deducidas mediante

el estudio de caminos en el grafo. Se define una noción de separación gráfica (d-separación) entre nodos, que está íntimamente ligada con la independencia de las variables aleatorias que estos representan.

En el Capítulo 4 se presenta el modelo de ecuaciones estructurales. Se definen los modelos intervenidos que permiten construir las variables contrafactuales, en el sentido definido en el Capítulo 2. Se enuncia el Teorema Back Door, que establece condiciones que indican como identificar la distribución de las variables contrafactuales. Por último, presentamos una manera alternativa para representar los sistemas intervenidos y una nueva demostración del Teorema de Back Door, utilizando las herramientas gráficas introducidas en el Capítulo 3, cuando la intervención se realiza en un único nodo.

La mayor parte de la bibliografía estudiada para la elaboración de este trabajo se encuentra citada a lo largo del mismo. También han resultado muy enriquecedoras las notas elaboradas por la Dra.Andrea Rotnitzky para el Curso "Inferencia Causal" en el X Congreso Monteiro [22].

El material que se encuentra en la lista bibliográfica pero que no fue citado fue también consultado para la confección de este trabajo.

A lo largo de los siguientes capítulos utilizaremos la siguiente notación: el símbolo $\coprod$ introducido por Dawid [3] denotará independencia e independencia condicional de variables aleatorias:

$$X \coprod Z \quad \text{denota la independencia entre las variables aleatorias } X \text{ y } Z \tag{1.0.1}$$

$$X \coprod Z \mid W \quad \text{denota la independencia entre } X \text{ y } Z \text{ condicional a la variable } W \tag{1.0.2}$$

Además si $(X, Y), (W, Z)$ son dos vectores aleatorios entonces

$$(X, Y) \sim (W, Z) \text{ significará que ambos vectores tiene la misma distribución.} \tag{1.0.3}$$

En muchos casos, las demostraciones presentadas se realzan para variables aleatorias discretas, con el próposito de enfatizar en los conceptos dejando de lado (importantes) tecnicismos.

Capítulo 2

Efecto medio del tratamiento

2.1. Presentación del problema

Siguiendo el enfoque propuesto por Hernán y Robins en algunos de sus trabajos ([9],[11],[12]), comenzaremos presentando el siguiente ejemplo para ilustrar el problema que se ha de abordar.

Supongamos que un paciente que estaba en espera para un trasplante de corazón, recibió el 1 de enero un corazón nuevo y cinco días más tarde murió. Imaginemos que de alguna manera podemos saber, tal vez por revelación divina, que si no hubiera recibido el corazón el 1 de enero (y todas las demás cuestiones de su vida se hubieran mantenido sin cambios) entonces hubiera estado vivo cinco dias después. La mayoría de las personas que cuentan con esta información, estarían de acuerdo en que el trasplante ha causado la muerte del paciente. La intervención tuvo un efecto causal en su sobrevida, cinco dias después.

Ahora pensemos que otra paciente recibió un trasplante el 1 de enero y cinco dias después está viva. Nuevamente imaginemos que podemos saber que si no hubiera recibido el corazón, igualmente seguiría viva. En esta paciente, el trasplante no tiene un efecto causal en su sobrevida. Esto ilustra como trabaja el razonamiento humano en la inferencia causal: comparamos (la mayoría de las veces mentalmente) el resultado cuando una acción está presente con el resultado cuando la acción está ausente, y el resto de los factores se mantienen inalterados. Si los dos resultados son diferentes decimos que la acción tuvo un efecto causal sobre el resultado o respuesta en el individuo; de lo contrario decimos que no observamos un efecto causal de la acción en el resultado de interés para ese individuo.

Identificar efectos causales individuales excede nuestras posibilidades en la medida en la que no podemos saber que es lo que hubiera ocurrido con un individuo si hubiera sido sometido a la acción contraria a la que en el se ejerció. Esta limitación demanda un enfoque diferente, dejando de lado la pregunta individual sobre el efecto causal de cierto tratamiento, para introducir el efecto medio del mismo en toda la población. Es decir, siguiendo con el ejemplo precedente, nos interesará comparar el porcentaje de sobrevida en hipotéticos escenarios donde (i)todos los pacientes sean transplantados o (ii)ningún paciente lo sea.

Reformulada la pregunta de interés científico (determinar el efecto medio del tratamiento), resta decidir si la información disponible (variables observadas) resulta suficiente para responder a esta nueva inquietud. En este sentido presentaremos las diferentes condiciones experimentales que pueden dar origen a los datos, y veremos bajo que condiciones podemos dar respuesta a la pregunta planteada.

2.2. Experimentos y estudio observacionales

Comenzaremos el abordaje de la problemática causal comentando brevemente la noción de dos importantes conceptos: experimentos (o diseños experimentales controlados) y estudios observacionales. Un experimento (ideal) es una investigación donde el sistema en estudio está bajo el control del investigador. Esto significa que tanto las personas o material investigado, la naturaleza de los tratamientos y la manera en la que estos son asignados, como así también las manipulaciones y los procedimientos de medición utilizados son seleccionados por el investigador.

En cambio, en un estudio observacional algunas de estas características, y en particular la asignación de los individuos a los grupos de diferentes tipos de tratamiento, se escapan del control del investigador.

El modelo contrafactual de causalidad es valioso precisamente porque ayuda a los investigadores a estipular supuestos, evaluar técnicas alternativas de análisis de datos, y pensar cuidadosamente sobre el proceso de exposición. Parte de su éxito se debe a la posibilidad que brinda al analista para conceptualizar los estudios observacionales como si fueran diseños experimentales controlados.

2.3. El modelo contrafactual o modelo causal de Rubin

El modelo causal de Rubin, también conocido como modelo de respuestas potenciales, consta de dos elementos fundamentales: las respuestas potenciales y el mecanismo de asignación del tratamiento.

Para poder desarrollar estos conceptos, siguiendo la notación introducida por Holland [14], denotemos con la letra U a la población que se pretende estudiar. Cada unidad-individuo en U es denotada por u. Para cada $u \in U$, hay asociado un valor $Y(u)$ de la variable de interés Y, a la que llamaremos respuesta observada. Además, se dispone de una segunda variable A definida en U cuyo valor en cada individuo indica a que acción este ha sido sometido. A modo de ejemplo, consideremos que cada individuo puede ser asignado a tratamiento o control. En tal caso, tendremos que $A(u) = t$ cuando el individuo u recibe tratamiento mientras que $A(u) = c$, caso contrario. Siguiendo el ejemplo presentado, podemos pensar que el tratamiento consiste en transplantarle al paciente un nuevo corazón. La manera en la que se determina el valor de A en cada individuo merece la siguiente definición.

Definición 2.3.1. Mecanismo de asignación es el método por el cual se determina la acción [1] a la que es sometido cada integrante de la población.

El par (A, Y) denota al conjunto de variables factuales (u observadas). En cada individuo u, $A(u)$ e $Y(u)$ indican el nivel de tratamiento al cual el individuo u fue sometido y la respuesta en él observada.

La idea clave para el desarrollo de la inferencia causal radica en la capacidad *potencial para exponer o no cada unidad a cierta acción y conceptualizar el valor de la variable respuesta de interés bajo cada uno de las posibles acciones.*

Es decir, cada unidad puede ser potencialmente expuesta a cualquiera de las posibles acciones. En nuestro ejemplo, las posibles acciones consisten en ser tratado o no (t y c, respectivamente). Los valores de la variable respuesta son potencialmente afectados por la acción, t o c, a la cual la unidad es expuesta. Necesitamos entonces introducir una variable respuesta para cada posible acción: $Y_t(u), Y_c(u)$.

Definición 2.3.2. Respuestas potenciales o contrafactuales. $Y_t(u)$, denota el valor de la respuesta que sería observada si la unidad u fuera expuesta a t mientras que $Y_c(u)$ es la respuesta que observaríamos si la unidad u fuera sometida al nivel c. Y_t, Y_c reciben el nombre de respuestas potenciales o contrafactuales.

[1] Llamaremos indistintamente: acción, tratamiento, exposición.

Esta presentación presupone que el valor de la variable respuesta en cada individuo sólo depende del tratamiento al cual éste fue sometido, independientemente de lo que ocurra con el nivel de tratamiento en los demás individuos de la población. Esta suposición se conoce en la literatura como **SUTVA** (stable-unit-treatment-value assumption) [24], y será asumida a lo largo del presente trabajo.

El efecto de t respecto a c en la unidad u medido en la variable respuesta de interés puede expresarse por

$$Y_t(u) - Y_c(u) \ . \tag{2.3.1}$$

Ahora bien, "*El problema fundamental de la inferencia causal"[14]* radica en la imposibilidad de observar los valores de Y_t y de Y_c en una *misma* unidad y, por consiguiente, no podemos observar efecto de t respecto a c en la unidad u, dado por la fórmula (2.3.1). La propuesta estadística para sortear esta dificultad consiste en estudiar el efecto medio (a lo largo de la población) de t respecto de c en la variable respuesta de interés, comparando

$$E[Y_t] \quad \text{con} \quad E[Y_c] \ .$$

Por ejemplo, podemos considerar como parámetro de interés causal el efecto medio del tratamiento (ATE: Average Treatment Effect), dado por

$$ATE = E[Y_t] - E[Y_c] \ . \tag{2.3.2}$$

El signo de esta diferencia indicaría que política adoptar a nivel poblaciónal. Siempre y cuando un valor mayor de Y indique un beneficio, $ATE > 0$ indica evidencias en favor del nivel de tratamiento t, mientras que $ATE = 0$ indica la falta de efecto medio del tratamiento a nivel poblaciónal. Surge así la siguiente definición.

Definición 2.3.3. Diremos que hay un efecto causal del tratamiento en la variable respuesta de interés si el efecto medio es diferente de cero:

$$E[Y_t] - E[Y_c] \neq 0 \ .$$

En general, un **parámetro causal** de interés es una cantidad que nos interesa conocer. El mismo se define en función de la pregunta del investigador. Nosotros, a modo de ejemplo, estudiaremos el efecto medio del tratamiento, dado por (2.3.2).

Típicamente, el parámetro causal de interés es un valor que depende de la distribución de las variables aleatorias contrafactuales. En este caso, la fórmula (2.3.2) depende de la distribución de las variables Y_t e Y_c. La variable respuesta factual Y (respuesta observada) se relaciona con las variables contrafactuales mediante la hipótesis de **consistencia**, que será asumida en lo que resta del trabajo.

$$Y(u) \ = \ \begin{cases} Y_t(u) & \text{si } A(u) = t, \\ Y_c(u) & \text{si } A(u) = c \ , \end{cases}$$

Es decir,

$$Y = I_{A=t} Y_t + I_{A=c} Y_c \ , \tag{2.3.3}$$

Entonces Y puede ser definida en términos de Y_t, Y_c y de A. Para profundizar en este concepto, se puede consultar la página 31 del trabajo de Hernan y Robins [12], como así también [1].

A modo de síntesis, consideremos la siguiente tabla (extraído de [18]), donde se ilustra como se relacionan las variables observadas con las variables contrafactuales en los diferentes grupos de tratamiento.

Grupo	Y_t	Y_c
Tratamiento$(A = t)$	observable como Y	contrafactual
Control $(A = c)$	contrafactual	observable como Y

Los efectos causales individuales se definen dentro de las filas de la tabla comparando Y_t e Y_c en cada individuo. Sin embargo, considerando que la variable Y_t es faltante en cada individuo del grupo control, resulta imposible el cálculo directo de los efectos causales a nivel individual sólo por medio de las variables observadas. Es por ello que el objeto de estudio serán parámetros causales poblaciónales, como por ejemplo el ATE. Surgen así los diferentes parámetros causales de interés.

Resta entonces determinar cuándo la distribución de las variables observadas (en nuestro ejemplo (A, Y)) determina el parámetro causal de interés. De esto trata la identificabilidad.

Definición 2.3.4. Identificabilidad: Si el parámetro causal de interés queda determinado mediante la distribución de las variables observadas, decimos que el mismo está identificado.

En los capitulos siguientes nos abocaremos a establecer condiciones que garanticen la identificabilidad. Para identificar el efecto medio causal, basta con que $E[Y_t]$ y $E[Y_c]$ queden determinadas a partir de la distribución de las variables observadas (o factuales). Además, es deseable proveer una fórmula que permita expresar el parámetro causal de interés a partir de la distribución de las variables observadas.

Estudiaremos en la próxima sección el problema de la indentificabilidad para el ejemplo que nos concierne: determinar bajo que condiciones $E[Y_t] - E[Y_c]$ queda determinado mediante la distribución de (A, Y).

Para finalizar esta sección, presentaremos algunas preguntas *causales* abordadas en la bibliografía mediante las herramientas que desarrollaremos en este trabajo.

Ejemplo 1: Escuelas públicas versus escuelas católicas. Consideraremos un ejemplo considerado por Morgan [17] en el cual se pretende determinar si los estudiantes del último año de escuela secundaria que asisten a la escuela católica en USA tienen mejor desempeño (puntuación) en la prueba de rendimiento estandarizada que los estudiantes del último año de escuela secundaria que asisten a la escuela pública. En este contexto, el modelo contrafactual presupone que los estudiantes tienen dos resultados posibles en la prueba de rendimiento: uno que se observaría si fueran educados en la escuela católica y otro que se observaría si fueran educados en escuela pública (vamos a suponer que en los efectos de aprendizaje todas las católicas son iguales y los de las públicas también).

Se define A la variable "tratamiento" siendo

$$A = \begin{cases} t & \text{si el estudiante asiste a escuela católica} \\ c & \text{si el estudiante asiste a escuela pública.} \end{cases}$$

La variable factual o respuesta observada está dada por

$$Y = \text{puntaje obtenido por el estudiante en la prueba de rendimiento.}$$

y se definen las respuestas (o variables) contrafactuales :

$$Y_t = \text{puntaje en la prueba de rendimiento del estudiante si asistiera a escuela católica,}$$

$$Y_c = \text{puntaje en la prueba de rendimiento del estudiante si asistiera a escuela pública.}$$

Recordemos que la hipótesis de consistencia establece que

$$Y = I_{A=t}Y_t + I_{A=c}Y_c.$$

Por lo tanto la distribución de la variable observada Y contiene sólo una parte de la información de las variables contrafactuales y por ello sin hipótesis adicionales no es evidente que podamos usar las variables observadas A e Y para hallar la distribución de Y_t y de Y_c.

En este ejemplo, $E[Y_t]$ representa la nota promedio en una situación hipotética en la que todos los estudiantes concurren a escuela católica, mientras que $E[Y_c]$ representa la nota promedio en una situación hipotética en la que todos los alumnos concurren a escuela pública. $ATE > 0$ sugeriría dejar la educación secundaria en manos de escuelas católicas, mientras que $ATE < 0$ sería una evidencia en favor de las escuelas públicas.

Ejemplo 2: Formación Laboral. Consideraremos el problema de estudiar el efecto causal de la formación laboral de recursos humanos sobre los ingresos futuros, presentado en Heckman et al [8]. Para ello, los posibles niveles de tratamiento están dados por

$$A = \begin{cases} t & \text{si el empleado participa en el programa de entrenamiento} \\ c & \text{caso contrario.} \end{cases} \tag{2.3.4}$$

La respuesta observada está dada por

$$Y = \text{salario anual del empleado}$$

mientras que las variables contrafactuales son:

$Y_t =$ ingreso anual en presencia del "tratamiento"(es decir, si es el empleado es entrenado),
$Y_c =$ ingreso anual en ausencia de entrenamiento.

En este ejemplo $E[Y_t]$ representa el valor promedio de ingresos anuales si todos empleados participaran del programa de capacitación mientras que $E[Y_c]$ es el valor promedio de ingresos anuales si ningún empleado fuera capacitado.

$ATE > 0$ estaría indicando que la capacitación genera un incremento en la media del salario anual de los empleados.

Ejemplo 3: Efecto de un fármaco [26]. Supongamos que se desea estudiar el efecto de cierto fármaco, según la dosis suministrada. A diferencia de los ejemplos tratados hasta el momento, en este caso para cada posible valor x de la dosis del fármaco consideramos una variable respuesta potencial: Y_x es la respuesta que un sujeto tendría si recibiera la dosis x.

Denotemos por X a la dosis asignada a cada paciente. La variable respuesta Y puede ser binaria (por ejemplo: si se alivia el dolor de cabeza a la hora de haber recibido la droga o no) o continua: escala (0 a 100) que mide el nivel del dolor de cabeza 1 hora despues de haber recibido el fármaco. Bajo el supuesto de consistencia, la respuesta Y coincide con el valor de la variable contrafactual correspondiente a la dosis asignada: si $X(u) = x$, entonces $Y(u) = Y_x(u)$. En otras palabras, tenemos que $Y(u) = Y_{X(u)}(u)$, para cada individuo $u \in U$. Resumiremos esta notación poniendo

$$Y = Y_X.$$

En este caso, las respuestas contrafactuales se convierten en un proceso contrafactual

$$\{(Y_x) : x \in \mathbb{R}_{\geq 0}\}.$$

Estudiar como varía $E[Y_x]$ en función de x resulta ser la pregunta causal de interés. Esta pregunta excede el alcance de este trabajo, pero no queríamos dejar de presentar el ejemplo, invitando al lector interesado a profundizar sus conocimientos en el tema.

Ejemplo 4: Trasplante de corazón. Supongamos que queremos saber el efecto causal de trasplantar el corazón en la población de pacientes con cierta disfunción cardíaca. Imaginemos que, una vez detectada la enfermedad, los pacientes pueden ser transplantados o no, en cuyo caso se los medicará. Nos referiremos al trasplante como exposición mientras que hablaremos de control en alusión al tratamiento farmacológico. Nos interesa estudiar la sobrevida dentro de los primeros seis meses, a partir del momento en que se determinó la acción a ser aplicada en el paciente. Vamos a considerar la variable dicotómica A, siguiendo con la notación de los ejemplos anteriores, siendo

$$A = \begin{cases} t & \text{si el paciente es trasplantado (expuesto, tratado)} \\ c & \text{si el paciente no es trasplantado} \end{cases} \tag{2.3.5}$$

y una variable de respuesta Y también dicotómica, dada por

$$Y = \begin{cases} 1 & \text{si el paciente muere dentro de los seis meses,} \\ 0 & \text{caso contrario.} \end{cases} \tag{2.3.6}$$

Las variables contrafactuales Y_t e Y_c toman valores en el conjunto $\{0, 1\}$ y satisfacen

$$Y_t = 1 \quad \text{si el paciente muere dentro de los seis meses habiendo sido transplantado (presencia del "tratamiento")} \tag{2.3.7}$$

$$Y_c = 1 \quad \text{si el paciente muere dentro de los seis meses sin haber sido transplantado.} \tag{2.3.8}$$

Definimos $P(Y_t = 1)$ como la proporción de sujetos que mueren dentro de los seis meses, si todos los individuos fueran trasplantados. Llamamos **riesgo de** Y_a, para $a = t, c$, a $P(Y_a = 1)$. Para variables Bernoulli, la exposición tiene un efecto causal en la población si

$$P(Y_t = 1) \neq P(Y_c = 1) \ ,$$

indicando diferencias en el riesgo para los diferentes tratamientos.

Observación 2.3.1. *Si la respuesta es binaria,* $ATE = P(Y_t = 1) - P(Y_c = 1)$

2.3.1. Falta de identificabilidad sin restricciones

Demostraremos, mediante un ejemplo, que si no se hacen supuestos sobre la distribución conjunta del tratamiento y las variables contrafactuales no es posible identificar $E[Y_t]$, $E[Y_c]$ y tampoco $ATE = E[Y_t] - E[Y_c]$, a partir de la distribución de las variables observadas (A, Y). Considerando variables discretas. Para demostrar este hecho, debemos encontrar funciones de probabilidad puntual para (Y_t^1, Y_c^1, A^1) y (Y_t^2, Y_c^2, A^2) de forma tal que

$$(Y^1, A^1) \sim (Y^2, A^2)$$

pero

$$ATE^1 := E[Y_t^1] - E[Y_c^1] \ \neq \ ATE^2 := E[Y_t^2] - E[Y_c^2] \ .$$

Notación 2.3.1. *recordando lo expuesto en (1.0.3)* $(Y^1, A^1) \sim (Y^2, A^2)$ *significará que ambos vectores tiene la misma distribución.*

Construiremos un ejemplo de este fenómeno donde Y será una variable dicotómica, tomando los valores 0 y 1, de forma tal que $E[Y_a^j] = P(Y_a^j = 1)$, para $j = 1, 2$, $a = t, c$. Consideremos la siguiente función de probabilidad puntual para el vector (Y_c^1, Y_t^1, A^1):

	(Y_c^1, Y_t^1)			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$A^1 = c$	0	1/4	1/4	0
$A^1 = t$	1/4	0	0	1/4

Haciendo uso de la hipótesis de consistencia, la función de probabilidad puntual asociada a las variables observadas (Y^1, A^1) está dada por

$A^1 \setminus Y^1$	0	1
c	1/4	1/4
t	1/4	1/4

donde, por ejemplo,

$$P(Y^1 = 1, A^1 = c) = P(Y_c^1 = 1, Y_t^1 = 0, A^1 = c) + P(Y_c^1 = 1, Y_t^1 = 1, A^1 = c) = 1/4 + 0 = 1/4$$

y en particular

$$\begin{aligned} P(Y_t^1 = 1) &= P(Y_t^1 = 1, A^1 = t) + P(Y_t^1 = 1, A^1 = c) \\ &= P(Y_c^1 = 0, Y_t^1 = 1, A^1 = t) + P(Y_c^1 = 1, Y_t^1 = 1, A^1 = t) \\ &+ P(Y_c^1 = 0, Y_t^1 = 1, A^1 = c) + P(Y_c^1 = 1, Y_t^1 = 1, A^1 = c) \\ &= 0 + 1/4 + 1/4 + 0 \\ &= 1/2. \end{aligned}$$

Ahora consideremos la siguiente función de probabilidad puntual para el vector (Y_c^2, Y_t^2, A^2)

	(Y_c^2, Y_t^2)			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$A^2 = c$	0	1/4	0	1/4
$A^2 = t$	0	1/4	1/4	0

Tenemos entonces que

$A^2 \setminus Y^2$	0	1
c	1/4	1/4
t	1/4	1/4

de donde concluímos que (Y^2, A^2) la misma distribución que (Y^1, A^1): $(Y^2, A^2) \sim (Y^1, A^1)$.

Sin embargo,

$$\begin{aligned} P(Y_t^2 = 1) &= P(Y_t^2, A^2 = t) + P(Y_t^2 = 1, A^2 = c) \\ &= P(Y_c^2 = 0, Y_t^2 = 1, A^2 = t) + P(Y_c^2 = 1, Y_t^2 = 1, A^2 = t) \\ &+ P(Y_c^2 = 0, Y_t^2 = 1, A^2 = c) + P(Y_c^2 = 1, Y_t^2 = 1, A^2 = c) \\ &= 1/4 + 0 + 1/4 + 1/4 \\ &= 3/4 \end{aligned}$$

con lo cual

$$P(Y_t^2 = 1) \neq P(Y_t^1 = 1).$$

Como nuestro objetivo es calcular el ATE, faltan calcular $P(Y_c^1 = 1)$ y $P(Y_c^2 = 1)$:

$$\begin{aligned} P(Y_c^1 = 1) &= P(Y_c^1 = 1, A^1 = t) + P(Y_c^1 = 1, A^1 = c) \\ &= P(Y_c^1 = 1, Y_t^1 = 1, A^1 = t) + P(Y_c^1 = 1, Y_t^1 = 0, A^1 = t) \\ &+ P(Y_c^1 = 1, Y_t^1 = 0, A^1 = c) + P(Y_c^1 = 1, Y_t^1 = 1, A^1 = c) \\ &= 1/4 + 0 + 1/4 + 0 \\ &= 1/2 \end{aligned}$$

y

$$\begin{aligned} P(Y_c^2 = 1) &= P(Y_c^2 = 1, A^2 = t) + P(Y_c^2 = 1, A^2 = c) \\ &= P(Y_c^2 = 1, Y_t^2 = 1, A^2 = t) + P(Y_c^2 = 1, Y_t^2 = 0, A^2 = t) \\ &+ P(Y_c^2 = 1, Y_t^2 = 0, A^2 = c) + P(Y_c^2 = 1, Y_t^2 = 1, A^2 = c) \\ &= 0 + 1/4 + 0 + 1/4 \\ &= 1/2 \end{aligned}$$

de esta forma $\begin{cases} ATE^1 = & P(Y_t^1 = 1) - P(Y_c^1 = 1) = 1/2 - 1/2 = 0 \\ ATE^2 = & P(Y_t^2 = 1) - P(Y_c^2 = 1)) = 3/4 - 1/2 = 1/4 \end{cases}$ con lo cual

$$ATE^1 \neq ATE^2 \ ,$$

tal como queríamos demostrar.

Como vemos, dos diferentes distribuciones contrafactuales generan la misma distribución de los datos observados. Los datos observados no nos permiten deducir (identificar) cual de las dos distribuciones de variables contrafactuales consideradas produjo los datos observados.

2.3.2. Asociación vs. Causalidad

El principal objetivo de esta sección es discutir la diferencia entre medidas de asociación $E[Y \mid A = t] - E[Y \mid A = c]$ y medidas de efecto causal, como la que presentamos en este trabajo: $E[Y_t] - E[Y_c]$. Queremos entender que representan

$$E[Y \mid A = t] - E[Y \mid A = c] \quad \text{y} \quad E[Y_t] - E[Y_c] \,. \tag{2.3.9}$$

Volviendo al ejemplo de la educación católica o pública, en el lado izquierdo de (2.3.9) estaríamos representando la diferencia de las medias de las puntuaciones observadas en las pruebas de rendimiento en dos subconjuntos disjuntos de la población: los que fueron a escuela católica y los que asistieron a escuela pública, mientras que el lado derecho representaría la diferencia media de las variables contrafactuales Y_a para $a = c, t$ en la población *entera* de interés.

Definición 2.3.5. Si $E[Y \mid A = t] - E[Y \mid A = c] \neq 0$ decimos que la respuesta está asociada al tratamiento.

En particular, si hay asociación, tenemos que $Y \mid A = t \not\sim Y \mid A = c$. La distribución de las variables observadas permite determinar si éstas están o no asociadas, sin necesidad de ningún tipo de supuesto. Sin embargo, nosotros queremos determinar causalidad, es decir, comparar los valores medios de las variables contrafactuales, como lo indica la Definición 2.3.3. Es por ello que queremos enfatizar en la diferencia existente entre asociación y causalidad. En la próxima Sección veremos bajo qué condiciones estos conceptos coinciden.

2.4. Identificabilidad I: Intercambiabilidad - Aleatorización

Aleatorización es un mecanismo de asignación de los tratamientos que garantiza que tratados y no tratados conforman grupos **intercambiables**. Para lograr esto, el mecanismo por el cual se conforman dichos grupos debe ser independiente del pronóstico de los individuos (o de los potenciales resultados).

En tal caso, toda variable aleatoria W se distribuye de igual forma entre tratados y no tratados:

$$W|A = c \sim W|A = t \,.$$

Para lograr esto, el mecanismo por el cual se conforman dichos grupos debe ser ajeno a los resultados de los experimentos a los que se someten los individuos. En particular, las respuestas contrafactuales tienen igual distribución en los grupos definidos por $A = c$ y $A = t$, y por consiguiente, resultan independientes del mecanismo de asignación del tratamiento. Surge entonces la siguiente definición.

Definición 2.4.1. Diremos que se verifican las condiciones de aleatorización si Y_a es independiente de A, para $a = t, c$. Siguiendo la notación introducida en (1.0.1)

$$Y_a \coprod A \,, a = t, c \,.$$

En tal caso, decimos que tratados y no tratados conforman grupos intercambiables.

Una posibilidad consiste en utilizar una urna con bolitas con los nombres de los individuos y mediante extracciones a ciegas elegimos quien recibe el tratamiento. Este tipo de mecanismos generan poblaciónes (tratados y controles) homogéneas, que resultan intercambiables. En tales circunstancias,

el grupo control respondería con la misma distribución que se observa en los tratados, caso ellos mismos lo fueran. Esto nos permite predecir el comportamiento que el grupo control tendría en caso de que hubiera sido tratado a través de los resultados observados en el grupo de tratado.

A modo de ejemplo, supongamos que tenemos una respuesta binaria, indicando éxito o fracaso. Sea N_t la cantidad de individuos bajo tratamiento y N_c la cantidad individuos en el grupo control. Denotemos con n_t la cantidad de individuos bajo tratamiento que tuvieron repuesta satisfactoria. La intercambiabilidad entre el grupo de tratados y el grupo control nos permite decir que que la proporción de individuos no tratados que hubieran obtenido respuesta satisfactoria en caso de que hubieran sido tratados coincide con la observada en los tratados y está dada por n_t/N_t. Podemos entonces concluir que

$$P(Y_t = 1) \;=\; \frac{n_t + \frac{n_t}{N_t} N_c}{N_t + N_c} \;=\; \frac{n_t}{N_t} \;=\; P(Y = 1 \mid A = t)\ .$$

Análogamente, si n_c denota la cantidad de individuos dentro del grupo control que tuvieron respuesta satisfactoria, tenemos que

$$P(Y_c = 1) \;=\; \frac{n_c + \frac{n_c}{N_c} N_t}{N_t + N_c} \;=\; \frac{n_c}{N_c} \;=\; P(Y = 1 \mid A = c)\ .$$

Veremos ahora como podemos generalizar esta idea a partir de la definición de aleatorización. En el ejemplo precedente, resulto fundamental garantizar la presencia de individuos bajo cada nivel de tratamiento que se pretende estudiar. Esta condición resultará fundamental para poder identificar la distribución de la correspondiente variable contrafactual. Surgen así las llamadas condiciones de **positividad**. En el presente contexto necesitamos individuos en el grupo tratamiento y en el grupo control. Para ello, se requiere que $0 < P(A = a) < 1,\ \ a = t, c$

Lema 2.4.1. *Si* $0 < P(A = t) < 1$*, bajo aleatorización, tenemos que*

$$E[Y_a] = E[Y \mid A = a] \quad , \quad a = t, c\ .$$

Demostración. Bajo las condiciones del Lema, tenemos que

$$E[Y \mid A = a] = E[Y_a \mid A = a] = E[Y_a]$$

La primer igualdad vale por consistencia (ecuación (2.3.3)) y la segunda vale por aleatorizacion. □

Corolario 2.4.1. *Bajo aleatorización, asociación es causalidad:*

$$E[Y_t] - E[Y_c] \;=\; E[Y|A = t] - E[Y|A = c]\ .$$

Ejemplos de experimentos en lo que se verifica la condicion de aleatorización son aquellos donde para cada individuo *se lanza una misma moneda* (no necesariamente equilibrada) para determinar si recibe el tratamiento $(A = t)$ o no $(A = c)$. En tal caso, la probabilidad de recibir tratamiento t es la misma para todos los individuos.

Observación 2.4.1. *En* (Y_t^1, Y_c^1, A^1) *de la Sección 2.3.1 vale la aleatorizacion, pues*

$$P(Y_t^1 = 1 \mid A = c) = P(Y_t^1 = 1 \mid A = t)$$
$$P(Y_c^1 = 1 \mid A = c) = P(Y_c^1 = 1 \mid A = t)$$

Observación 2.4.2. *En* (Y_t^2, Y_c^2, A^2) *de la Sección 2.3.1 NO vale la aleatorizacion, pues*

$$P(Y_c^2 = 1 \mid A = c) \neq P(Y_c^2 = 1 \mid A = t)$$

2.5. Identificabilidad II: Aleatorización Condicional

Consideremos un diseño experimental en el que la asignación de los individuos a los grupos se realiza considerando una covariable [2] L (puede ser un vector) que podemos medir en todos los individuos y que, en cada nivel de la covariable $L = \ell$, tratados y no tratados resultan intercambiables. En tales circunstancias, hablamos de aleatorización condicional.

Volviendo al ejemplo de los trasplantes de corazón introducido en la página (8), supongamos que los investigadores tienen medida la variable *pronóstico* L, que establece si los pacientes se encuentran o no en condición crítica, según $L = 1$ o $L = 0$, respectivamente ($L = 1 : crítica, L = 0 : no\ \ crítica$). Una vez medida esta covariable, podemos diferenciar dos grupos: los que tienen condición crítica y los que no, es decir los de mal pronóstico y los de mejor. La variable L ocurre antes de la asignación del tratamiento. Supongamos ahora que para cada nivel $L = \ell$, se lanza una moneda para conformar los grupos tratamiento-control, siendo que la probabilidad con la que la moneda asigna al grupo tratamiento depende de ℓ. La ayuda de la moneda permite asumir que, para cada nivel $L = \ell$, tratados y no tratados conforman grupos intercambiables y, por consiguiente, toda variable aleatoria tiene la misma distribución en estos dos grupos. En particular, Y_t e Y_c satisfacen esta propiedad y por consiguiente

$$Y_a|A = t, L = \ell \sim Y_a|A = c, L = \ell\ ,\ \text{para } a = t, c.$$

Definición 2.5.1. Diremos que se verifican las condiciones de aleatorizacion condicional si Y_a es independiente de A dada la variable L, para $a = t, c$.

$$Y_a \coprod A \mid L.$$

Resta garantizar la presencia de individuos bajo el nivel de tratamiento que se pretende estudiar, en cada nivel de la covariable L: determinar las condiciones de **positividad**. En el presente contexto, para poder identificar $E[Y_c]$ necesitamos garantizar que en cada nivel de $L = \ell$ existan personas que fueron asignadas al grupo control: $P(A = c|L = \ell) >$, siempre que $P(L = \ell) > 0$.

El supuesto de intercambiabilidad condicional garantiza identificabilidad.

Lema 2.5.1. *Supongamos que la variable respuesta, A y L son discretas. Si $P(A = a|L = \ell) > 0$ cada vez que $P(L = \ell) > 0$, bajo aleatorización condicional y consistencia, tenemos que*

$$P[Y_a = y] = \sum_{\ell} P[Y = y \mid A = a, L = \ell].P\,[L = \ell] \tag{2.5.1}$$

y por consiguiente

$$E[Y_a] = \sum_{y,\ell} y\, P[Y = y \mid A = a, L = \ell].P\,[L = \ell]$$

queda determinada a partir de la distribución de las variables observadas. En particular, cuando la respuesta es binaria, tenemos que

$$E[Y_a] = P[Y_a = 1] = \sum_{\ell} P[Y = 1 \mid A = a, L = \ell].P\,[L = \ell]\ . \tag{2.5.2}$$

[2] Referiremos a una *covariable* L como una variable aleatoria medible que no es ni tratamiento A ni respuesta$Y$.

Demostración: $P[Y_a = y] = \sum_{\ell} P[Y_a = y, L = \ell]$

$$= \sum_{\ell} P[Y_a = y \mid L = \ell].P[L = \ell]$$
$$= \sum_{\ell} P[Y_a = y \mid A = a, L = \ell].P[L = \ell] \text{ aleatorización condicional}$$
$$= \sum_{\ell} P[Y = y \mid A = a, L = \ell].P[L = \ell] \text{ consistencia}$$

□

Observemos que $\sum_{\ell} P[Y_a = 1, L = \ell] = \sum_{P(L=\ell)>0} P[Y_a = y, L = \ell]$ y que para que (2.5.2) valga, es necesario que $P[A = a, L = \ell] > 0$ cada vez que $P(L = \ell) > 0$.

El resultado precedente admite la siguiente generalización:

Lema 2.5.2. *(Estandarización) Si $P(A = a|L = \ell) > 0$ cada vez que $P(L = \ell) > 0$, bajo aleatorización condicional y consistemcia, tenemos que*

$$E[Y_a] = E[E[Y|A = a, L]] .$$

Demostración: $E[Y_a] = E[E[Y_a|L]]$ propiedad
$= E[E[Y_a|A = a, L]]$ aleatorización condicional
$= E[E[Y|A = a, L]]$ consistencia

Para justificar la segunda igualdad es pertinente observar que bajo aleatorización condicional

$$Y_a \coprod A \mid L = \ell \Rightarrow Y_a| \ L = \ell \sim Y_a| \ L = \ell, A = a .$$

□

Corolario 2.5.1. *Bajo aleatorización condicional , si $0 < P(A = t|L = \ell) < 1$ cada vez que $P(L = \ell) > 0$, tenemos que $E[Y_t] - E[Y_c]$ queda determinado por la distribución de las variables observadas (L, A, Y):*

$$E[Y_t] - E[Y_c] = E[E[Y|A = t, L]] \text{ -}E[E[Y|A = c, L]]$$

Corolario 2.5.2. *En particular si la respuesta es binaria y L y A son discretas se tiene:*

$$E[Y_t] - E[Y_c] = \sum_{l} P[Y = 1 \mid A = t, L = \ell].P[L = \ell] - \sum_{l} P[Y = 1 \mid A = c, L = \ell].P[L = \ell] .$$

Lema 2.5.3. *(Ponderación con probabilidad inversa) Bajo aleatorización condicional, consistencia y $P(A = a \mid L = \ell) > 0$ tenemos que*

$$E[Y_a] = E\left[\frac{I_{A=a}Y}{P(A = a \mid L)}\right].$$

Demostración:

$$
\begin{aligned}
E[Y_a] &= E\left[\frac{P(A=a \mid L)Y_a}{P(A=a \mid L)}\right] \\
&= E\left[\frac{P(A=a \mid Y_a, L)Y_a}{P(A=a \mid L)}\right] \text{ independencia condicional} \\
&= E\left[E(I_{A=a} \mid Y_a, L)\frac{Y_a}{P(A=a \mid L)}\right] \text{ esperanza = probabilidad de variable binaria} \\
&= E\left[E\left(I_{A=a}\frac{Y_a}{P(A=a \mid L)} \mid Y_a, L\right)\right] \text{ propiedades de esperanza condicional} \\
&= E\left[\frac{I_{A=a}Y_a}{P(A=a \mid L)}\right] \text{ esperanza de la esperanza} \\
&= E\left[\frac{I_{A=a}Y}{P(A=a \mid L)}\right] \text{ consistencia}
\end{aligned}
$$

□

Volvamos a un caso de población finita.

Supongamos nuevamente que tenemos una respuesta binaria, indicando éxito o fracaso. Consideremos las siguientes cantidades:

N : cantidad de personas en la población
N_ℓ : cantidad de personas en el nivel $L = \ell$
$N_{\ell a}$: cantidad de personas en el nivel $L = \ell$ a la que se le aplica el nivel de tratamiento a
$n_{\ell a}$: cantidad de personas del nivel $L = \ell$ y tratamiento a y que tuvieron respuesta positiva (éxito).

La intercambiabilidad condicional entre el grupo de tratados y el grupo control a nivel $L = \ell$ nos permite decir que la proporción de individuos no tratados que hubieran obtenido respuesta satisfactoria en caso de que hubieran sido tratados coincide con la observada en los tratados, a cada nivel $L = \ell$ es decir, y está dada por $n_{\ell t}/N_{\ell t}$. Esto nos permite determina qué pasaría en la población si todos los individuos fueran tratados. Para ello, combinaremos los resultados observados en aquellos que efectivamente han sido tratados e, intercambiabilidad condicional de por medio, predeciremos el comportamiento que se hubiera observado en los no tratados, en caso de que hubieran sido tratados. Más específicamente, entre los individuos con $L = \ell$, $N_{\ell,t}$ han sido tratados y en $n_{\ell,t}$ el resultados fue positivo. La intercambiabilidad condicional indica que $(n_{\ell,t}/N_{\ell,t})\,\%$ de los $N_{\ell c}$ hubieran tenido respuesta positiva, en caso de que hubieran sido tratados. Por lo tanto, entre los individuos con $L = \ell$, tenemos que la cantidad de éxitos que se observarían en caso de que toda la población fuera tratada está dada por

$$n_{\ell,t} \; + \frac{n_{\ell,t}}{N_{\ell t}} N_{\ell c} \; .$$

Sumando a lo largo de todos los niveles de la variable L, concluímos que

$$P(Y_t = 1) \;=\; \frac{1}{N} \sum_\ell \left\{ n_{\ell t} + \frac{n_{\ell t}}{N_{\ell t}} N_{\ell c} \right\} \;=\; \frac{1}{N} \sum_\ell \frac{n_{\ell t}}{N_{\ell t}} N_\ell = \frac{1}{N} \sum_\ell \frac{n_{\ell t}}{N_{\ell t}/N_\ell} \tag{2.5.3}$$

La cantidad $\frac{N_{\ell a}}{N_l} = P(A = a \mid L = \ell)$, para $a = c, t$, se conoce como Propensity Score, y utilizaremos $\pi_a(\ell)$ para denotar $P(A = a \mid L = \ell)$. Por lo tanto,

$$P(Y_t = 1) = \frac{1}{N} \sum_{\ell} \frac{n_{\ell t}}{\pi_t(\ell)} \ .$$

Es decir, a cada individuo de la población con respuesta favorable y $L = \ell$ que ha recibido el nivel de tratamiento t, se le asigna un peso inversamente proporcional al Propensity Score correspondiente a su nivel ℓ.

Notemos también que

$$P(Y_t = 1) = \sum_{\ell} \frac{n_{\ell t}}{N_{\ell t}} \frac{N_\ell}{N} = \sum_{\ell} P(Y = 1|L = \ell, A = t)P(L = \ell)$$

Obteniéndose nuevamente la fórmula dada en el Lema 2.5.1

2.6. Estudios observacionales

A diferencia de lo que ocurre en los *experimentos*, donde tratados y no tratados han sido asignados según un diseño definido por el investigador (aleatorización, aleatorización condicional), en los estudios observacionales el investigador no controla el mecanismo de asignación del tratamiento. En tales circunstancias, para poder identificar el parámetro causal de interés, hipótesis adicionales acerca de la distribución de las variables contrafactuales son requeridas. A modo de ejemplo hemos visto en la sección 2.3.1 que la no identificabilidad de los efectos causales en estudios observacionales proviene del hecho que la distribución de los datos observados es consistente con diferentes valores del parámetro de interés.

Típicamente, en estudios observacionales se trata de encontrar un conjunto de variables L de forma tal que resulte razonable suponer que tratados y no tratados son condicionalmente intercambiables en cada nivel de L. En tales circunstancias, supondremos que vale la aleatorización condicional y procederemos a analizar los datos como si hubieran sido obtenidos siguiendo tal diseño. Los resultados obtenidos están sujetos a la validez de esta suposición.

Cabe enfatizar que las hipótesis requeridas para identificar no son testeables a partir de la distribución de los datos observados. A modo de ejemplo, hemos viso que la distribución presentada en el ejemplo de la Sección 2.3.1 para las variables observadas podría provenir de (A^1, Y_t^1, Y_c^1), donde vale la aleatorización, o bien de (A^2, Y_t^2, Y_c^2), donde la aleatorización no es verificada. Este mismo tipo de dificultad sufre la condición de aleatorización condicional. Es por ello que resulta fundamental decidir con los expertos cual es el conjunto de variables L a ser considerado para que la hipótesis de independencia condicional resulte razonable y valga la identificabilidad. En tales circunstancias, decimos que no hay variables confusoras no medidas. No hay confusión. Hay identificabilidad.

Decimos que hay confusión cuando otras variables, además de las incluídas en L, deben ser medidas y condicionadas para lograr intercambiabilidad condicional.

2.7. El ejemplo

Profundizaremos ahora el Ejemplo 4, introducido en la Sección 2.3 del presente Capítulo, para ilustrar los conceptos desarrollados en las últimas Secciones. Seguiremos los trabajos de Hernán [9] Hernán y Robins[11]

Calcularemos efectos causales bajo tres posibles escenarios. En los primeros dos asumiremos que los datos han sido obtenidos siguiendo un diseño de aleatorización y aleatorización condicional, respectivamente. Por último, asumiremos que los datos provienen de un estudio observacional, estudiaremos el

supuesto de aleatorización condicional para poder identificar el parámetro causal de interés y discutiremos por qué los supuestos que hacemos pueden ser controversiales en tales circunstancias.

Recordemos que en el Ejemplo 4, interesaba estudiar el efecto causal del trasplante de corazón en la sobrevida de los pacientes con cierta disfunción cardíaca. Las variables A e Y fueron definidas en (2.3.5) y (2.3.6), y representan el tratamiento asignado al paciente (t =trasplantado) y el resultado obtenido: $Y = 1$ cuando el paciente muere antes de los seis meses. Las variables Y_t e Y_c toman el valor 1 si el paciente muere dentro de los seis meses, en el caso en que fuera trasplantado (Y_t) o no (Y_c) (ver ecuaciones (2.3.7) y (2.3.8), respectivamente). Recordemos que la hipótesis de consistencia vincula estas variables mediante la fórmula $Y = I_{A=t}Y_t + I_{A=c}Y_c$.

Definiremos $P(Y_a = 1)$ como el **Riesgo** de morir cuando $a = t, c$, y lo llamaremos riesgo contrafactual. También definiremos $P(Y = 1 \mid A = a)$ que será el riesgo de morir observado entre los que recibieron el tratamiento a, para $a = t, c$.

En este contexto una posible medida del efecto causal de interés es la llamada "diferencia de riesgo causal" que, en presencia de respuestas binarias, coincide con el Average Treatment Effect: $P(Y_t = 1) - P(Y_c = 1) = E[Y_t] - E[Y_c]$.

Diseño 1: Aleatorización no condicional. Consideremos la tabla 2.7.1

Id	A	Y
a	c	0
b	c	1
c	c	0
d	c	0
e	t	0
f	t	0
g	t	0
h	t	1
i	c	1
j	c	1
k	c	0
l	t	1
m	t	1
n	t	1
o	t	1
p	t	1
q	t	1
r	t	0
s	t	0
t	t	0

Tabla 2.7.1: Incluye los datos observados:A,Y

.

Supongamos que estos valores han sido obtenidos bajo **aleatorización**, es decir, los individuos a ser transplantados han sido seleccionados al azar. Bajo este diseño los expuestos y no expuestos son intercambiables y, por consiguiente, el riesgo de mortalidad contrafactual bajo cada valor de exposición (t ó c) es el mismo en los expuestos que en los no expuestos:

$$P(Y_a = 1 \mid A = t) = P(Y_a = 1 \mid A = c)\,,$$

y por consiguiente,

$$P(Y_a = 1) = P(Y = 1 \mid A = a) \quad \text{para } a = c, t \ .$$

Considerando que estamos asumiendo que los datos presentados en la tabla se obtuvieron siguiendo un diseño de experimento aleatorizado, vale al intercambiabilidad y, como vimos en la Sección 2.4, el riesgo contrafactual $P(Y_a = 1)$ a nivel de exposicion a, para $a = t, c$, es igual al riesgo observado entre los que recibieron ese mismo nivel de exposición: $P(Y_a = 1) = P(Y = 1 \mid A = a)$, para $a = t, c$.

La diferencia de riesgo observado entre los que recibieron diferentes niveles de exposición $P(Y = 1 \mid A = t) - P(Y = 1 \mid A = c)$ es calculada de los datos disponibles para el par (A, Y). De hecho, observando la Tabla 2.7.1 , tenemos que $P(Y = 1 \mid A = t) - P(Y = 1 \mid A = c) = 7/13 - 3/7$.

En síntesis, en un experimento realizado bajo estas condiciones, $Y_a \amalg A$ asegura que se puede medir el efecto causal medio del tratamiento a partir de la distribución del par (A, Y) y por consiguiente, ATE resulta identificable a partir de la distribución de las variables observadas. No hay *confusión* .

Diseño 2: Aleatorización condicional. Imaginemos ahora que los datos presentados en la tabla 2.7.1 provienen de la tabla 2.7.2, donde se incluye el valor de la variable pronóstico L en cada individuo, siendo que la variable L toma el valor 1 si el individuo se encuentra en condición crítica.

$$L \ = \ \text{Pronóstico} \ = \ \begin{cases} 1 & \text{si el sujeto tiene condición crítica,} \\ 0 & \text{caso contrario.} \end{cases} \tag{2.7.1}$$

Id	L	A	Y
a	0	c	0
b	0	c	1
c	0	c	0
d	0	c	0
e	0	t	0
f	0	t	0
g	0	t	0
h	0	t	1
i	1	c	1
j	1	c	1
k	1	c	0
l	1	t	1
m	1	t	1
n	1	t	1
o	1	t	1
p	1	t	1
q	1	t	1
r	1	t	0
s	1	t	0
t	1	t	0

Tabla 2.7.2: Incluye la variable pronóstico L.

Notemos que la distribución de L difiere entre tratados y no tratados. De hecho, $P(L = 1|A = c) = 3/7$ mientras que $P(L = 1|A = t) = 9/13$. Este hecho indica que tratados y no tratados NO son intercambiables. Por consiguiente, el supuesto de aleatorización no es correcto y entonces el Diseño 1 (aleatorización) no ha sido aplicado.

Sin embargo, imaginemos que el investigador nos informa que en realidad la asignación de pacientes a los grupos se realizó aleatorizando en cada nivel de la variable L. Es decir, los médicos clasificaron a todas las personas en condición crítica, y no crítica, y seleccionaron al azar 75 % (9 de 12) ($P(A = t|L = 1) = 9/12$) de las personas críticas y 50 %(4 de 8) de las no críticas para ser trasplantadas. Estaríamos entonces en presencia de datos generados bajo **aleatorización condicional.** (Diseño 2).

En este caso, usando las fórmula presentadas en el Lema 2.5.1, concluímos que $P(Y_t = 1) - P(Y_c = 1) = 0$.

Observemos que un estudio con aleatorización condicional puede ser considerado como una combinación de dos experimentos aleatorizados por separados: uno proveniente del subconjunto de personas en estado crítico ($L = 1$) y el otro del subconjunto de personas en estado no crítico ($L = 0$).

Paradigma de los experimentos aleatorizados para estudios observacionales

Consideremos ahora un estudio en el cual los investigadores no intervienen en la asignación de trasplante. Podemos entonces pensar que la Tabla 2.7.2 contiene los datos que ellos recogieron, como ocurre en los estudios observacionales. La Tabla indica que A no ha sido aleatorizado ya que la variable L se distribuye de manera diferente estre tratados y no tratados. Por lo tanto

$$Y_a \not\perp\!\!\!\perp A,\ a = t, c$$

puesto que la variable L se distribuye de manera diferente entre tratados y no tratados. Sin embargo, los expertos que decidieron a quien trasplantar, consideran razonable suponer que vale la aleatorización condicional, condicionando en la variable pronóstico L. Es decir, si bien el tratamiento no ha sido asignado según un protocolo donde se garantiza la aleatorización condicional, consideran que la manera en que el tratamiento fue asignado podría ser asumida como tal. A modo de ejemplo, si L fuese la única variable observada por los médicos al momento de asignar el tratamiento y el 50 % de médicos trasplanta cuando $L = 0$, mientras que el 75 % lo hace cuando $L = 1$, no habría objeción en asumir que vale la aleatorización condicional, con $P(A = t|L = 0) = 1/2$ mientras que $P(A = t|L = 1) = 9/12$

En los estudios observacionales *esta decisión puede ser considerada practicamente un acto de fe.* La aleatorización condicional es una suposición que no puede ser refutada ni avalada a partir de los datos disponibles. Asumirla es una determinación que se toma en forma conjunta con los expertos del área, entendiendo que es lo que ésta significa. Tenemos entonces que un estudio observacional, como el que estamos considerando, puede ser visto como un experimento realizado bajo aleatorización condicional en el cual

la intercambiabilidad condicional no está garantizada pero es asumida con la ayuda del conocimiento de los expertos en el área.

Si la hipótesis de los investigadores acerca de la intercambiabildad condicional es correcta entonces el riesgo causal puede ser identificado usando el método de estandarización (lema 2.5.2) o el de ponderación con probabilidad inversa (lema 2.5.3).

Como ya se enfatizó, los investigadores no pueden chequear la hipótesis de aleatorización condicional porque la respuesta contrafactual Y_a no es completamente observada. Como resultado de esta imposibildad, resulta controversial la inferencia causal en estudios observacionales.

2.8. Perdemos la identificabilidad por condicionar de más

En ocasiones al analizar un estudio observacional el investigador condiciona en todas las variables disponibles, como un modo de controlar potenciales variables de confusión. Sin embargo, es fundamental elegir apropiadamente el vector L para que la suposición de independencia condicional resulte razonable.

Agregar variables dentro del vector L no tiene por qué *ayudar* a obtener la independencia condicional. Daremos ahora un ejemplo numérico donde

1. Vale la aleatorización: las variables contrafactuales son independientes de la asignación del tratamiento: $Y_a \coprod A$, para $a = t, c$ y por consiguiente tenemos que $E[Y_a] = P(Y = 1|A = a)$.

2. Disponemos de una variables extra L y decidimos utilizarla para estandarizar y calcular ATE siguiendo las fórmulas presentadas en 2.5.1, llegando a un resultado equivocado.

Este ejemplo pretende ilustrar que condicionar en variables que no son confusoras puede conducir a conclusiones erradas. De aquí la importancia del conocimiento brindado por los especialistas a la hora de elegir las covariables que vamos a medir para resolver nuestro problema de identificación.

Construyamos un ejemplo de esta situación, especificando la distribución conjunta de covariables, variables contrafactuales y tratamiento: (A, Y_t, Y_c, L). Cabe enfatizar que este tipo de información nunca está a disposición del investigador, quien únicamente puede aspirar a conocer la distribución conjunta de las variables observadas (L, A, Y). Consideremos la siguiente Tabla de probabilidad conjunta de (A, Y_t, Y_c, L)

	(Y_t, Y_c, L)							
	$(0,0,0)$	$(0,1,0)$	$(0,0,1)$	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(0,1,1)$	$(1,1,1)$
$A = t$	0	0	0	0	0	1/9	2/9	0
$A = c$	0	4/9	0	0	0	0	0	2/9

Tabla 2.8.1: Probabilidades Puntuales de (A,Y_t,Y_c,L).

A partir de la Tabla 2.8.1 podemos obtener la distribución de las variables contrafactuales Y_c e Y_t, de donde deducimos que

$$ATE = P(Y_t = 1) - P(Y_c = 1) = 1/3 - 1.$$

Tenemos además que la función de probabilidad puntual del vector (L, A, Y) está dada por

	(Y, L)			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$A = t$	0	2/9	1/9	0
$A = c$	0	0	4/9	2/9

Tabla 2.8.2: Probablidades conjuntas de las variables observadas.

Marginalizando obtenemos que la función de probabilidad conjunta de (A, Y_t, Y_c) está dada por:

	(Y_t, Y_c)			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$A = t$	0	2/9	0	1/9
$A = c$	0	4/9	0	2/9

Tabla 2.8.3: Probabilidades puntuales de Tratamiento y Contrafactuales.

Tenemos entonces que se verifica la condición de aleatorización: $Y_a \coprod A$ para $a = t, c$. Utilizando la tabla observada 2.8.2, constatamos que

$$ATE = P(Y_t = 1) - P(Y_c = 1) = P(Y = 1|A = t) - P(Y = 1|A = c) \, .$$

Por otra parte, tenemos que

	(Y_t, L)			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$A = t$	0	2/9	1/9	0
$A = c$	4/9	0	0	2/9

Notemos que $P(Y_t = 1|A = t, L = 0) \neq P(Y_t = 1|A = c, L = 0)$ (el lado izquierdo es 1 y el derecho es 0), y $P(Y_t = 1|A = t, L = 1) \neq P(Y_t = 1|A = c, L = 1)$ (el lado izquierdo es 0 y el lado derecho es 1), y por consiguiente, no se verifica la aleatorización condicional respecto de la variable L. Sin embargo, si cometiéramos el error de asumir esta condición y utilizarámos la fórmula presentada en el Corolario 2.5.1 con la información disponible en la tabla observada, obtenemos que

$$\sum_{\ell} P(Y = 1|A = t, L = \ell)P(L = \ell) - \sum_{\ell} P(Y = 1|A = c, L = \ell)P(L = \ell) = 5/9 - 1.$$

El valor obtenido no coincide con ATE, pues $1/3 - 1 \neq 5/9 - 1$.

Conclusión 2.8.1. *En el ejemplo que consideramos, las respuestas contrafactuales son independientes del nivel de tratamiento:* $Y_a \coprod A$, *para* $a = t, c$ *y por consiguiente* $P(Y_t = 1) - P(Y_c = 1) = P(Y = 1|A = t) - P(Y = 1|A = c)$. *Sin embargo, el hecho de condicionar (o ajustar) en la variable* L, *NO nos proporciona independencia condicional entre el tratamiento y las repuestas contrafactuales.Utilizar las fórmulas propuestas para identificar asumiendo independencia condicional puede conducir a conclusiones erradas.*

Para finalizar la presente Sección, queremos mencionar que con la Tabla observada (2.8.2) podemos verificar las suposiciones relacionadas con la positividad: $0 < P(A = t) < 1$ o bien $0 < P(A = t|L = \ell) < 1$, cuando $P(L = \ell)$, según trabajemos bajo aleatorización o aleatorización condicional, respectivamente. Sin embargo, no podemos verificar el supuesto de aleatorización ni el de aleatorización condicional. Es por ello que en estudios observacionales, donde se desconoce el mecanismo de asignación del tratamiento pero se necesita asumir algún tipo de aleatorización para garantizar la identificabilidad, la determinación del vector L debe realizarse con la ayuda del experto puesto que no son los datos quienes permiten dilucidar este tipo de inquietudes. Estudiaremos en el próximo Capítulo una propuesta gráfica para saber en qué variables debemos condicionar (o ajustar).

2.9. Cotas

¿Qué sucede si no podemos identificar el parámetro causal de interés a partir de la distribución de las variables observadas? En tal caso procuraremos dar cotas para el mismo. Volviendo al ejemplo que estamos considerando en este trabajo, si mayores valores de la variable respuesta indican mejor condición, cuando la cota inferior que obtenemos para ATE es positiva, tendremos evidencias a favor del tratamiento, por más que no podamos determinar el valor exacto de ATE. Mas aún, podemos garantizar que hay un efecto medio del tratamiento, según la Definición 2.3.3.

A continuación mostraremos como podemos acotar el efecto medio del tratamiento. Haremos un abordaje algo sofisticado para el problema que queremos tratar con la intención de ilustrar cuáles son las herramientas con las que se pueden trabajar en otras situaciones.

Teorema 2.9.1. *Sea Y una variable con distribución Bernoulli. Sea q la distribución de las variables observadas (A, Y), de forma tal que*

$A\backslash Y$	0	1
t	q_1	q_2
c	q_3	q_4

Tenemos entonces que

$$-q_1 - q_4 \leq E[Y_t] - E[Y_c] \leq q_2 + q_3 \ .$$

Demostración: Denotemos con $\mathcal{M}_k$ al simplex de dimensión k:

$$\mathcal{M}_k = \left\{ x = (x_1, x_2, \ldots, x_k) : x_i \geq 0, \sum_{i=1}^{k} x_i = 1 \right\}.$$

La variable A toma valores en el conjunto $\{c, t\}$ mientras que las respuestas contrafactuales Y_t, Y_c toman valores en $\{0, 1\}$. Cada posible función de probabilidad puntual para (A, Y_t, Y_c) puede identificarse con un elemento en $\mathcal{M}_8$. Para fijar notación, consideremos la Tabla 2.9.1

	(Y_t, Y_c)			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$A = t$	x_1	x_2	x_3	x_4
$A = c$	x_5	x_6	x_7	x_8

Tabla 2.9.1: Tabla de probabilidad puntual de A y las variables contrafactuales

Sea L_{mg} la aplicación que a cada elemento $x \in \mathcal{M}_8$ le asigna la función de probabilidad puntual del par (A, Y), siendo $Y = Y_t I_{A=t} + Y_c I_{A=c}$ cuando $(A, Y_t, Y_c) \sim x$

$$L_{mg} : \mathcal{M}_8 \longrightarrow \mathcal{M}_4$$
$$L_{mg}(x) \sim (A, Y), \text{ si } (A, Y_t, Y_c) \sim x$$

de forma que la distribución puntual de las variables observadas está dada por la Tabla 2.9.2

$A\backslash Y$	0	1
t	$x_1 + x_2$	$x_3 + x_4$
c	$x_5 + x_7$	$x_6 + x_8$

Tabla 2.9.2: Probabilidad puntual observada

Consideremos ahora la aplicación

$$L_{caus} : \mathcal{M}_8 \longrightarrow \mathbb{R}$$
$$L_{caus}(x) = E[Y_t] - E[Y_c], \text{ si } (A, Y_t, Y_c) \sim x$$

Con los datos de la Tabla 2.9.1 podemos calcular las probabilidades $P(Y_t = 1), P(Y_c = 1)$. Por ejemplo

$$P(Y_t = 1) = \left.\begin{array}{ccccc} P(A = t, Y_t = 1, Y_c = 0) & + & P(A = t, Y_t = 1, Y_c = 1) \\ + P(A = c, Y_t = 1, Y_c = 0) & + & P(A = c, Y_t = 1, Y_c = 1) \end{array}\right\} = x_3 + x_4 + x_7 + x_8$$

De la misma forma calculamos $P(Y_c = 1) = x_6 + x_8 + x_2 + x_4$ y con esto ya tenemos una fórmula para $L_{caus}(x) = x_3 + x_7 - x_2 - x_6$. Consideremos $q \in \mathcal{M}_4$, la función de probabilidad puntual para las variables observadas

$A\backslash Y$	0	1
t	q_1	q_2
c	q_3	q_4

Calculemos

$$m(q) = mín_{x \in \mathcal{M}_8 : L_{mg}(x) = q} L_{caus}(x)$$
$$M(q) = máx_{x \in \mathcal{M}_8 : L_{mg}(x) = q} L_{caus}(x)$$

Empecemos calculando el máximo, queremos hacer máximo la siguiente expresión $x_3 + x_7 - x_2 - x_6$

con las siguientes restricciones $\left\{\begin{array}{ccccc} x_1 & + & x_2 & = & q_1 \\ x_3 & + & x_4 & = & q_2 \\ x_5 & + & x_7 & = & q_3 \\ x_6 & + & x_8 & = & q_4 \end{array}\right.$.

Podemos hacer $x_2 = x_6 = 0$ que es el valor menor que pueden tomar. Ahora si pensamos en hacer max los valores de x_3 y x_7, de acuerdo con las restricciones necesitamos $x_7 = q_3$ y $x_3 = q_2$, con lo cual

$$M(q) = máx_{x \in \mathcal{M}_8 : L_{mg}(x) = q} L_{caus}(x) = q_2 + q_3.$$

Con la misma idea ahora buscamos obtener el mínimo de $x_3 + x_7 - x_2 - x_6$, sujeto a las condiciones ya mencionadas, obteniendose $x_2 = q_1$ y $x_6 = q_4$, con lo cual

$$m(q) = mín_{x \in \mathcal{M}_8 : L_{mg}(x) = q} L_{caus}(x) = -q_1 - q_4.$$

De esta forma podemos acotar como se deseaba:

$$m(q) \leq L_{caus}(x) \leq M(q) \ , \forall x : L_{mg}(x) = q$$
$$-q_1 - q_4 \leq L_{caus}(x) \leq q_2 + q_3 \ , \forall x : L_{mg}(x) = q \ .$$

□

Capítulo 3

DAG's

Las funciones de probabilidad pueden ser asociadas a grafos de forma tal que condiciones de independencia o independencia condicional, como las requeridas para garantizar la identificabilidad de parámetros causales, pueden ser verificadas utilizando herramientas de esta teoría.

3.1. Grafos: Algunas definiciones

Un *grafo* $G = (V, E)$ se define por medio de un conjunto V finito de *vértices o nodos* y un conjunto $E \subseteq V \times V$ de *aristas* que conectan los vértices. En nuestras aplicaciones los vértices representarán variables aleatorias y las aristas indicarán relaciones entre estas. Dos vértices que se conectan por una arista serán llamados *adyacentes*. Cuando $(u, v) \in E$ pero $(v, u) \notin E$ escribimos $u \to v$ (ó $v \leftarrow u$) y decimos que la arista es dirigida o que hay una *flecha (dirigida)* de u a v. En este caso diremos que u es *padre* de v y v es *hijo* de u. El conjunto de padres de u se nota $pa_G(v)$. Además, si $W \subseteq V$ se defne el conjunto de padres de W siendo

$$pa_G(W) \ = \ \cup_{v \in W} pa_G(v). \tag{3.1.1}$$

Definición 3.1.1. Sea $G = (V, E)$ un grafo. Si todas las aristas en E son dirigidas diremos que G es un *grafo dirigido*.

Un *camino* es una sucesión de nodos adyacentes. Por ejemplo, un *camino* de v a w está dado por $(v_1, v_2), (v_2, v_3), (v_3, v_4)...(v_{n-1}, v_n)$, con $v_1 = v$, $v_= w$, siendo que (v_i, v_{i+1}) o (v_{i+1}, v_i) es una arista en el grafo. Por lo general, indicamos el camino mediante el conjunto de vértices que este une y lo denotaremos poniendo $v \leftrightarrow w$.

Un *camino dirigido* es una sucesión de flechas dirigidas, de forma tal que cada una empieza con el vértice con el cual termina la flecha precedente. Por ejemplo, un *camino dirigido* de v a w está dado por $(v_1, v_2), (v_2, v_3), (v_3, v_4)...(v_{n-1}, v_n)$, con $v_1 = v$, $v_= w$, de forma tal que (v_i, v_{i+1}) es una flecha dirigida en el grafo. Por lo general, indicamos el camino mediante el conjunto de vértices que este une: $v_1, v_2, v_3, v_4, \ldots, v_{n-1}, v_n$. La existencia de un camino dirigido de v a w se denota con $v \mapsto w$ y decimos que v es un *antecesor o ancestro* de w, mientras que w se dice un *descendiente* de v. El conjunto de antecesores de v lo escribimos $an_G(v)$, al de descendientes $de_G(v)$. Estas definiciones también se extienden a conjuntos de nodos, tomando uniones, como en (3.1.1) Por ejemplo,

$$an_G(Z) \ = \ \cup_{z \in Z} an_G(z). \tag{3.1.2}$$

Notemos que

$$v \in an_G(W) \ \Leftrightarrow de_G(v) \cap W \neq \emptyset \ .$$

Definición 3.1.2. Diremos que un grafo dirigido $G = (V, E)$ es acíclico si para todo $v \in V$ no existe camino dirigido de v a v. En tal caso, diremos que G en un *grafo dirigido acíclico* (DAG: directed acyclic graph).

3.2. Distribuciones compatibles con un DAG G - La factorización Markov

Sea N el cardinal de V. Etiquetemos el conjunto de nodos de manera compatible con el grafo G: $V = \{v_1, v_2, \ldots, v_N\}$ donde

$$an_G(v_i) \subset \{v_1, \ldots, v_{i-1}\} .$$

Definición 3.2.1. Sea $(X_{v_1}, \ldots, X_{v_N})$ un vector aleatorio, indexado mediante los nodos de G, con función de probabilidad conjunta P. Diremos que la distribución P es compatible con el grafo G si la distribución de la variable X_{v_j} condicional a $\{X_{v_i} : i \leq j\}$ coincide con la distribución que se obtiene al condicionar en las variables correspondientes a los padres del nodo v_j:

$$X_{v_j} \mid \{X_{v_i}, i \leq j-1\} \sim X_{v_j} \mid \{X_{v_i} : v_i \in pa_G(v_j)\} . \tag{3.2.1}$$

Cuando resulte conveniente, supondremos que las variables aleatorias son discretas, para poder enfatizar los conceptos y dejar de lado los formalismos requeridos para generalizar estas ideas. De hecho, cuando las variables son discretas, la condición 3.2.1 puede ser escrita como

$$P(X_{v_j} = x_j | X_{v_i} = x_j, i \leq j-1) = P\left(X_{v_j} = x_j | X_{v_i} = x_i, i : v_i \in pa_G(v_j)\right) \tag{3.2.2}$$

siempre que las probabilidades condicionales estén bien definidas.

Ejemplo 3.2.1. *Sea G el grafo*

$$v_1 \longrightarrow v_2 \longrightarrow \cdots \longrightarrow v_{N-1} \longrightarrow v_N$$

Dada la definición precedente, tenemos que la distribución del vector $(X_{v_1}, \ldots, X_{v_N})$ es compatible con G si para $i \geq 2$ se verifica que

$$P\left(X_{v_i} = x_i | X_{v_j} = x_j, j \leq i\right) = P\left(X_{v_i} = x_i | X_{v_{i-1}} = x_{i-1}\right) . \tag{3.2.3}$$

Es decir, las distribuciones compatibles con G son aquellas correspondientes a cadenas de Markov finitas.

En adelante utilizaremos $P(x_{i_1}, \ldots, x_{i_k})$ para denotar $P(X_{v_{i_1}} = x_{i_1}, \ldots, X_{v_{i_k}} = x_{i_k})$. Mas generalmente, utilizaremos $P(x_i | \, x_s, s \in S)$ para $P(X_{v_i} = x_i | X_{v_s} = x_s, s \in S)$, siendo $S \subset \{1, \ldots, N\}$.

La *regla multiplicativa de la probabilidad* nos permite descomponer a P como un producto de N distribuciones condicionales, siempre que éstas estén bien definidas, según la siguiente fórmula

$$P(x_1, x_2, \ldots, x_N) = \prod_j P(x_j \mid x_1, x_2, \ldots, x_{j-1}).$$

Si la distribución P es compatible con el grafo G, cada uno de estos factores coincide con

$$P(x_j \mid x_1, x_2, \ldots, x_{j-1}) = P(x_j \mid x_i : v_i \in pa_G(v_j))$$

Queda entonces probado el siguiente resultado.

Lema 3.2.1. *La distribución P del vector aleatorio $(X_{v_1}, \ldots, X_{v_N})$ es compatible con el grafo G si admite la siguiente descomposición*

$$P(x_1, x_2, \ldots, x_N) = \prod_{j=1}^{N} P\left(x_j \mid x_i : v_i \in pa_G(v_j)\right) \ . \tag{3.2.4}$$

Ejemplo 3.2.2. *El siguiente DAG con nodos* $\{v_1, v_2, v_3, v_4, v_5\}$

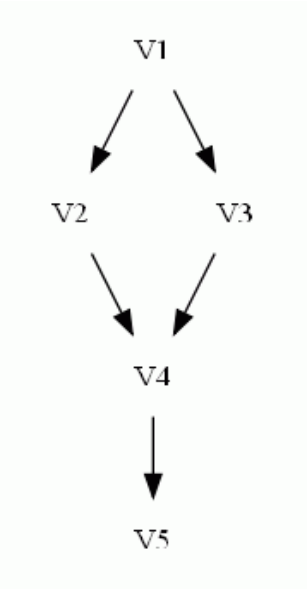

Figura 3.2.1: Los nodos V_i estan asociados a las variables X_{V_i}

induce la descomposición $P(x_1, x_2,, x_3, x_4, x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(x_5 \mid x_4)$

Definición 3.2.2. *Compatibilidad Markov:* Si una función de probabilidad P admite una descomposición como en 3.2.4 con respecto al DAG G, decimos que G *representa* a P, que G y P son compatibles o que P es Markov relativo a G. En tal caso, diremos que (3.2.4) es la descomposición markoviana de P.

Daremos a continuación una forma sistemática de construir distribuciones compatibles con un grafo G. Para comenzar, notemos que en el libro *Acoplamiento e processos Estocásticos,* Ferrari y Galves [5] definen a $(X_1, X_2, \ldots)$ siendo una cadena de Markov con espacio de estados E si existe una función $F : E \times [0,1] \to E$ tal que para todo $n \geq 2$

$$X_i = F(X_{i-1}, U_i) \ ,$$

con $U_1, U_2, \ldots,$ variables aleatorias independientes con distribución $U_i \sim \mathcal{U}[0,1]$. En tal caso, tenemos que

$$P\left(X_i = x_i | X_j = x_j, j \leq i-1\right) = P(X_i = x_i | X_{i-1} = x_{i-1}) = P(F(x_{i-1}, U_i) = x_i) \ . \tag{3.2.5}$$

Tenemos entonces que $(X_1, X_2, \ldots)$ es además una cadena de Markov homogénea. De hecho, si utilizamos funciones F_i se preserva la propiedad Markoviana: sea $F_i : E \times [0,1] \to E$ y $(X_1, X_2, \ldots)$ satisfaciendo

$$X_i = F_i(X_{i-1}, U_i) \ ,$$

con $U_1, U_2, \ldots$, i.i.d., $U_i \sim \mathcal{U}[0,1]$. Luego $(X_1, X_2, \ldots)$ satisface la propiedad (3.2.5) con F_i en lugar de F. Acabamos de dar una manera de construir distribuciones compatibles con el DAG presentado en el Ejemplo 3.2.1. El siguiente resultado prueba que toda distribución compatible con un DAG G puede ser construída de esta manera.

Teorema 3.2.1. *Sea $(X_{v_1}, \ldots, X_{v_N}) \sim P$, siendo P una distribución compatible con G. Entonces, existen funciones $H_i(u, x_j : v_j \in pa_G(v_i))$, con $0 \leq u \leq 1$, y variables $U_i \sim \mathcal{U}[0,1]$, $1 \leq i \leq n$, independientes de forma tal que las variables $\hat{X}_i$ definidas mediante la recurrencia*

$$\hat{X}_{v_i} := H_i(U_i, \hat{X}_{v_j} : v_j \in pa_G(v_i)) \,, \tag{3.2.6}$$

forman un vector con distribución P: $(\hat{X}_1, \ldots, \hat{X}_N) \sim P$. Mas aún, todo vector construido mediante una recurrencia de la forma (3.2.6) y variables $\{U_i : v_i \in V\}$ independientes, tiene una distribución compatible con G.

Demostración: Sea $\mathcal{R}_i \subset \mathbb{R}$ el espacio donde toma valores la variable X_{v_i} (en el ejemplo de cadenas de Markov (3.2.1), $\mathcal{R}_i = E$ para todo i). Las funciones H_i pueden ser consideradas

$$H_i : [0,1] \times \bigotimes_{v_j \in pa_G(v_i)} \mathcal{R}_j \to \mathcal{R}_i \,.$$

Para construir las funciones H_i, consideremos la factorización de Markov dada en (3.2.4). Para cada i, para cada $\{x_j : v_j \in pa_G(v_i)\}$ con $P(\cap_{v_j \in pa_G(v_i)} X_{v_j} = x_j) > 0$, tenemos una distribución $P_i(\cdot \mid x_j : v_j \in pa_G(v_i))$ en el espacio $\mathcal{R}_i$ definida por

$$P_i\left(x \mid x_j : v_j \in pa_G(v_i)\right) = P\left(X_{v_i} = x \mid X_{v_j} = x_j : v_j \in pa_G(v_i)\right) \,.$$

Denotemos por $H_i(u, x_j : v_j \in pa_G(v_i))$ a la función inversa generalizada de esta distribución, de forma tal que si $U \sim \mathcal{U}[0,1]$,

$$H_i(U, x_j : v_j \in pa_G(v_i)) \sim P_i(\,\cdot \mid x_j : v_j \in pa_G(v_i)) \,. \tag{3.2.7}$$

Sean $\{U_i : 1 \leq i \leq N\}$ i.i.d., $U_i \sim \mathcal{U}[0,1]$. Utilizando las funciones $\{H_i : 1 \leq i \leq N\}$ construimos en forma recursiva variables aleatorias $\tilde{X}_i$ de la siguiente manera

$$\hat{X}_i = H_i(U_i, \hat{X}_j : v_j \in pa_G(v_i)) \,.$$

Las independencia de las variables U_i y (3.2.7) garantizan que el vector $(\hat{X}_1, \ldots, \hat{X}_N)$ tiene misma distribución que $(X_1, \ldots, X_N)$:

$$\begin{aligned} P\left(\hat{X}_i = x_i \mid \hat{X}_j = x_j : j \leq n-1\right) &= P\left(H_i(U_i, x_j : v_j \in pa_G(v_i)) = x_i \mid \hat{X}_j = x_j : j \leq n-1\right) = \\ &= P\left(H_i(U_i, x_j : v_j \in pa_G(v_i)) = x_i\right) = P_i(x_i \mid x_j : v_j \in pa_G(v_i)) = \\ &= P\left(X_{v_i} = x_i \mid X_{v_j} = x_j : v_j \in pa_G(v_i)\right) \,. \end{aligned}$$

□

3.3. Representacion DAG de una distribución

Sea $(X_1, X_2, X_3, ..., X_N)$ un vector aleatorio con función de probabilidad P. Vamos a construir un grafo G_P de forma tal que la distribución P resulte compatible con G_P. Por cada variable X_i pondremos un nodo v_i. Para construir el conjunto de flechas, estudiaremos la factorización de P. Notemos que

$$P(x_1, x_2, \ldots, x_n) = \prod_j P(x_j \mid x_1, x_2, \ldots, x_{j-1}).$$

Supongamos ahora que la distribución de la variable X_j condicional a sus antecesores $(X_1, \ldots, X_{j-1})$, depende sólo de un subconjunto de ellos, que denotaremos con $PA_j \subset \{X_1, \ldots, X_{j-1}\}$ y denominaremos *padres markovianos* de X_j. Pondremos una arista dirigida en G_P entre v_i y v_j si $X_i \in PA_j$. Tautológicamente, tenemos P es compatible con G_P. El grafo resultante es un DAG y el par (G_p, P) se llama *red bayesiana.*

3.4. Métodos gráficos para estudiar independencias condicionales

De cierta forma, podemos pensar que un DAG es un modelo probabilístico indicando las condiciones de independencia que las distribuciones deben satisfacer para pertenecer al modelo. Una forma posible de caracterizar el conjunto de distribuciones compatibles con un DAG G es listar las independencias (también condicionales) que cada distribución debe satisfacer. Estas independencias se pueden 'leer' en el DAG utilizando un criterio gráfico llamado *d-separación* [19] (d denota la dirección). Este criterio se utiliza para conocer qué relaciones de independencia condicional son verificadas por las distribuciones compatibles con el grafo G, estudiando algunos de sus caminos. Para poder presentar los resultados existentes en este sentido, necesitaremos introducir algunas definiciones relativas a estructuras que pueden estar presentes en los caminos (no necesariamente dirigidos) de un grafo.

Configuraciones de los DAG's

Definición 3.4.1. Sea $G = (V, E)$ un grafo acíclico dirigido. Consideremos un camino p, no necesariamente dirigido.

- Cadena: diremos que el camino p tiene una cadena si incluye la siguiente estructura:

$$v_i \longrightarrow v_j \longrightarrow v_k$$

- Tenedor: diremos que el camino p tiene un tenedor con centro en v_s si incluya la siguiente estructura:

$$v_r \longleftarrow v_s \longrightarrow v_t$$

- Colisionador: diremos que el camino p tiene una colisionador (o tenedor invertido) en v_g si incluya la siguiente estructura:

$$v_e \longrightarrow v_g \longleftarrow v_f$$

Definición 3.4.2. (d-separación) Un camino p en un DAG G se dirá *bloqueado* por un conjunto de nodos $Z = \{z_1, \ldots, z_k\} \subset V$ si se verifica al menos una de las siguientes condiciones:

- p contiene una cadena o un tenedor con centro en Z:
  cadena: $v_i \longrightarrow z_j \longrightarrow v_k$, con $z_j \in Z$
  tenedor: $v_i \longleftarrow z_j \longrightarrow v_k$, con $z_j \in Z$
- p contiene un colisionador de forma tal que ni él ni sus descendientes pertenecen a Z:
  $v_e \longrightarrow v_g \longleftarrow v_f$, con $v_g \notin Z$ y $de(v_g) \cap Z = \emptyset$.

Cuando $Z = \emptyset$ decimos que p está bloqueado si tiene un colisionador.

Definición 3.4.3. Dado un DAG $G = (V, E)$, consideremos tres subconjuntos disjuntos de nodos W, T y Z. Diremos que Z *d-separa* a los conjuntos W y T si Z bloquea *todos* los caminos p que unen un vértice de W con uno de T. Cuando $Z = \emptyset$, decimos que W y T están d-separados si todo camino p que une un nodo de W con un nodo de T tiene un colisionador.

Notación 3.4.1. *$(W \coprod T \mid Z)_G$ significa que W y T están d-separados por Z en el grafo G. Dado un vector aleatorio $(X_v : v \in V)$ y un subconjunto $W \subset V$, denotamos con X_W al subvector cuyas coordenadas pertenecen a W: $X_W = (X_w : w \in W)$. Si $X = (X_v : v \in V)$ tiene distribución P, la independencia entre X_W y X_T será denotada por $(X_W \coprod X_T)_P$, mientras que la independencia entre X_W y X_T condicional a X_Z se denotará mediante $(X_W \coprod X_T \mid X_Z)_P$.*

El próximo resultado permite caracterizar cuando son independientes subvectores de un vector cuya distribución P es compatible con un grafo G. La independencia de los subvectores está garantizada por la d-separación de los respectivos nodos en el grafo.

Teorema 3.4.1. *Sea $G = (V, E)$ un DAG, W y T dos subconjuntos disjuntos de nodos en V. Sea $(X_{v_1}, \ldots, X_{v_n})$ un vector aleatorio con distribución P compatible con G. Tenemos entonces que si $(W \coprod T)_G$, entonces $(X_W \coprod X_T)_P$, para toda P compatible con G.*

Demostración: Por el Teorema 3.2.1 sabemos que existen funciones $\{H_i : 1 \leq i \leq N\}$, y variables independientes U_i de forma tal que las variables definidas por la recursion

$$\hat{X}_i = H_i(U_i, \hat{X}_j : v_j \in pa_G(v_i))$$

conforman un vector con distribución P. Veamos entonces que $\hat{X}_W$ es independiente de $\hat{X}_T$. Para ello, notemos que $\hat{X}_T$ es función de las variables U_T y de $\{U_i : v_i \in an_G(T)\}$ mientras que $\hat{X}_W$ es función de U_W y de $\{U_i : v_i \in an_G(W)\}$, y por consiguiente, basta garantizar que $\{an_G(T) \cup T\}$ y $\{an_G(W) \cup W\}$ son conjuntos disjuntos para tener la independencia deseada. Este hecho se deduce de la d-separación de los conjuntos T y W en el grafo G. Afirmamos que

$$\{an_G(T) \cup T\} \cap \{an(W)_G \cup W\} \ = \emptyset \ .$$

Siendo

$$\{an_G(T) \cup T\} \cap \{an_G(W) \cup W\} = \{T \cap W\} \cup \{an_G(T) \cap W\} \cup \{an_G(W) \cap T\} \cup \{an_G(T) \cap an_G(W)\}$$

veamos que cada una de las intersecciones es vacía.

1. $\{T \cap W\} = \emptyset$ pues T, W eran disjuntos, por hipótesis.

2. $\{an_G(W) \cap T\} = \emptyset$ pues si suponemos que existe un elemento en la intersección, entonces $\exists j$: $t_j \in an_G(W) \Rightarrow \exists$ un camino dirigido entre t_j y algún w_i y esto es absurdo pues W y T están d-separados, entonces ningún camino esta compuesto sólo por cadenas.

3. $\{an_G(T) \cap W\} = \emptyset$ por simetría con $\{an_G(W) \cap T\} = \emptyset$.

4. $\{an_G(T) \cap an_G(W)\} = \emptyset$ porque si suponemos que existe un elemento en la intersección, entonces $\exists\ z \in an_G(T)$ y $z \in an_G(W)$, con lo cual hay un camino de un w_i a un t_j que tiene cadenas y un tenedor en z, contradiciendo la d-separación entre T y W.

□

Esquemáticamente, tendríamos

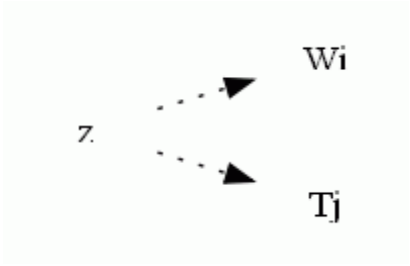

Figura 3.4.1: Las aristas punteadas significan caminos entre los $\text{an}_G(W)$ y $\text{an}_G(T)$

En realidad se puede demostrar que vale la recíproca del resultado anterior. Es decir, la independencia para toda distribución compatible con el DAG implica la d-separación. Mas aún, el siguiente teorema establece que la d-separación codifica todos las independencias condicionales lógicamente implícitas en la factorización de Markov de cualquier P compatible con un DAG G. Una demostración del siguiente resultado puede verse en Verma & Pearl [25], o Geiger.[6].

Teorema 3.4.2. *Sea $G = (V, E)$ un DAG, W, T y Z tres subconjuntos disjuntos de nodos en V. Sea $(X_{v_1}, \ldots, X_{v_n})$ un vector aleatorio con distribución P compatible con G. Tenemos entonces que*

$$(W \coprod T \mid Z)_G \ \Leftrightarrow\ (X_W \coprod X_T \mid X_Z)_P \text{ para toda } P \text{ compatible con } G\ .$$

Una de las ventajas de la utilización de redes bayesianas en el contexto causal, radica en la capacidad para representar y responder a los cambios externos o intervenciones, mediante modificaciones en la topología de la red, como veremos en el próximo Capítulo. Además, utilizaremos resultados presentados en el Teorema 3.4.1 y en el Teorema 3.4.2 para verificar las condiciones de identificabilidad requeridas en el Capítulo 2, en lo referente a independencia o independencia condicional.

Capítulo 4

Modelo de ecuaciones estructurales (SEM)

4.1. Ecuaciones estructurales

Consideremos un conjunto con n variables aleatorias $X = \{X_1, X_2, ..., X_n\}$, algunas de las cuales quizás no tenemos capacidad de medir. Supongamos que cada variable X_j está determinada por:

- un conjunto conocido de variables $PA_j \subseteq X - \{X_j\}$
- otra variable U_j, llamadas error, perturbacion o factor omitido, que no está determinada por X_j,

de forma tal que (PA_j, U_j) y X_j se relacionan determinísticamente por cierta función f_j, de la siguiente manera

$$X_j = f_j(PA_j, U_j) \ . \tag{4.1.1}$$

La ecuación 4.1.1 se la denomina *ecuación estructural*.

4.1.1. Diagramas causales

Dado un vector $X = (X_1, X_2, ..., X_n)$, satisfaciendo un sistema de de ecuaciones estructurales, podemos construir un grafo asociado al sistema de la siguiente manera.

Definición 4.1.1. Sea $X = (X_1, X_2, \ldots, X_n)$, satisfaciendo un sistema de ecuaciones estructurales, como en (4.1.1). Un *diagrama causal* será un grafo con un nodo v_j por cada variable X_j y pondremos una arista de v_i a v_j siempre que la variable X_i pertenezca al conjunto de variables PA_j. Además pondremos aristas punteadas bidirigidas entre todos los pares (X_j, X_k) si sus respectivos errores (U_j, U_k) no son independientes.

Observación 4.1.1. *Un diagrama causal es una representación gráfica del sistema de ecuaciones. No asume nada con respecto a las funciones f_i ni a la distribución de las perturbaciones U_i.*

Definición 4.1.2. Un diagrama causal en el que (a) no hay aristas punteadas bidirigidas, (b) el grafo resultante es acíclico y dirigido y (c) toda variable que es un determinante común de otras dos variables está incluida en el conjunto X como variable del sistema se dice un DAG causal.

4.2. Modelo de ecuaciones estructurales no paramétricas

Un sistema de ecuaciones estructurales asociado a un Grafo Causal se llama modelo de ecuaciones estructurales no paramétricas(NPSEM). El modelo está dado por el conjunto de funciones y la condición de independencia entre las coordenadas del vector $U = (U_1, \ldots, U_n)$. Para ser más específicos, consideremos la siguiente definición.

Definición 4.2.1. Un modelo causal M de ecuaciones estructurales no paramétricas (NPSEM) para las variables $X = (X_1, \cdots, X_n)$ asume que estas satisfacen un sistema de ecuaciones estructurales con funciones $\{f_i : 1 \leq i \leq n\}$, de forma tal que el diagrama causal asociado al sistema resulta un DAG causal. Es decir, se asume la existencia de:

1. un vector aleatorio $U = (U_1, U_2, ..., U_n)$ con coordenadas independientes, donde cada U_j es llamado error o perturbación
2. un conjunto de funciones desconocidas determinísticas f_j, de forma tal que las variables $\{X_1, \ldots, X_n\}$ quedan definidas mediante la siguiente recursión:

$$X_j = f_j(X_{s_1}, \ldots, X_{s_j}, U_j) \tag{4.2.1}$$

de forma tal que en el DAG causal G asociado a $\{f_i : 1 \leq i \leq n\}$, tenemos que

$$pa_G(v_j) = \{v_{s_1} \ldots, v_{s_j}\} .$$

En tal caso, decimos que las funciones $\{f_i : v_i \in V\}$ son compatibles con G.

Denotemos con $\mathcal{V}_i$ al conjunto donde toman valores las variables aleatorias asociadas al nodo v_i y utilizaremos $\mathcal{U}_i$ para el conjunto de posibles valores de U_i.Tenemos entonces que si $pa_G(v_j) = \{v_{s_1} \ldots, v_{s_j}\}$,para cada x_{s_i} en $\mathcal{V}_{s_i}$, u_j en $\mathcal{U}_j$, $f_j(x_{s_1}, x_{s_2,\ldots}, x_{s_j}, u_j) \in \mathcal{V}_j$. Cada una de las funciones f_j que aparece en la ecuación (4.2.1) representa un mecanismo por el cual se determina el valor de las variables de la izquierda (salida u output) a partir de los valores de las variables de la derecha: las variables precedentes $\{X_{s_1}, \ldots, X_{s_j} : v_{s_i} \in pa_G(v_j)\}$ y la perturbación U_j. La ausencia de una variable en la parte derecha de una ecuación codifica el supuesto de que la "naturaleza" omite esa variable en el proceso de determinación del valor de la variable de salida. Esta propiedad es de enorme utilidad y jugará un papel clave en el contexto causal a la hora de determinar las variables factuales y contrafactuales, como veremos en la Sección 4.4.

Observación 4.2.1. *Un modelo M de ecuaciones estructurales no paramétricas tiene asociado un DAG causal, y viceversa. Tenemos entonces la siguiente correspondencia*

$$G = (V, E) \qquad \Longleftrightarrow \qquad M = \begin{cases} U = \{U_1, \ldots, U_n\} & \textit{independientes,} \\ \{f_i : v_i \in V\} & \textit{compatible con } G\ , \end{cases}$$

4.3. Acerca de la notación

Vamos a dedicar algunas líneas a una cuestión menor pero no por ello menos importante. Se trata de la notación que los diferentes autores emplean a la hora de trabajar con grafos y vectores aleatorios. Nosotros, hasta el momento, hemos utilizado v_i para denotar en forma genérica los vértices de un grafo

G. Es decir, utilizamos $V = \{v_1, \ldots, v_n\}$ para denotar el conjunto de vértices de un grafo $G = (V, E)$. Por otra parte, los vectores aleatorios con los que hemos estado trabajando han sido denotados con $X = (X_1, \ldots, X_n)$, sabiendo que la coordenada X_i del vector está asociada con el nodo v_i del grafo. En adelante resultará de suma importancia enfatizar el carćter de la variable asociada al nodo. Por ejemplo, en el Capítulo 2, L representa una variable pre tratamiento, A la variable que indica el tratamiento recibido por cada individuo mientras que Y representa la respuesta observada en los individuos. La variable A juega un papel diferente, en el sentido de ser factible a ser manipulada o intervenida. Es decir, existen variables en las que podemos intervenir y otras que tan solo podemos observar. En el contexto causal, esta diferencia es fundamental, ya que interesa la distribución de las variable de respuesta bajo diferentes intervenciones (la distribución o algún parámetro asociado a la distribución de las variables cotrafactuales Y_a).

Los autores clásicos en el área utilizan la misma letra para denotar al nodo y a la variable aleatoria (factual) asociada. Cuando sea conveniente, adoptaremos esta notación, con las aclaraciones necesarias para evitar confusiones, como haremos en el siguiente ejemplo.

Ejemplo 4.3.1. *El sistema de ecuaciones estructurales*

$$\begin{aligned} Z &= f_Z(U_Z) \\ X &= f_X(Z, U_X) \\ Y &= f_Y(X, U_Y) \end{aligned}$$

tiene asociado el DAG causal

$$Z \to X \to Y.$$

La ausencia de la variable Z en los argumentos de f_Y transmite la afirmación de que variaciones en Z dejará a Y sin cambios, siempre y cuando las variables U_Y y X se mantengan constantes. Es decir, cada función es invariante a posibles cambios en la forma de las otras funciones.

4.4. Modelos intervenidos

Como empezamos explicando en la Sección anterior, un supuesto clave que queda implícito en los sistemas de ecuaciones estructurales es que la modificación de una de las funciones (pensemos en la correspondiente al nodo v_j), altera los valores de entradas de las ecuaciones correspondientes a los nodos descendientes de v_j, pero no la forma de las funciones restantes.

Podemos pensar en un circuito eléctrico complejo con cajas negras, donde cada ecuación representa un mecanismo aislado de forma tal que la j-ésima caja recibe como entrada a (PA_j, U_j) y devuelve X_j, habiendo operado según f_j. Si intervenimos y reemplazamos una de las cajas negras por alguna otra (cambiamos la ecuación asociada al nodo), estaríamos alterando el la salida de la misma y por consiguiente, la entrada de las cajas conectadas con ella. Es decir no se estaría alterando ninguna ecuación (mecanismos) que dicta el valor de la restantes variables, ni los valores de los errores (ya que están determinados por factores fuera del sistema).

Recordemos que las variables contrafactuales representan respuestas en escenarios hipotéticos donde se ha fijado el nivel de tratamiento a ser asignado en toda la población. Estas variables serán construídas mediante los NPSEM intervenidos, donde reemplazamos las funciones que crean las variables asociadas a la asignacion del tratamiento las constantes con la que pretendemos intervenir. En el NPSEM intervenido, las funciones correspondientes a los vértices de intervención no dependen del valor de ninguna otra variables del sistema, mientras que las demás funciones coinciden con las del sistema original. Es por

ello que el DAG asociado al NPSEM intervenido es idéntico al DAG original, salvo por el hecho de que todas las flechas que llegan a los vértices correspondientes a las variables intervenidas son eliminadas.

Definición 4.4.1. Dado un modelo causal M (Definición 4.2.1), un subconjunto de variables $A = \{A_1, \ldots, A_l\} \subset X$ asociado con los nodos $v_{i_1}, \ldots, v_{i_l}$ y un posible valor $\overline{a} = (a_{i_1}, \ldots, a_{i_l})$ para el vector A, el modelo intervenido $M_{\overline{a}}$, se define por:

1. el mismo vector $U = \{U_1, \ldots, U_n\}$ de perturbaciones que en el modelo M
2. un nuevo conjunto de ecuaciones, que coincide con las del modelo M excepto en lo que respecta a los nodos correspondientes a las variables en las que queremos intervenir. Es decir, si $\{f_i : v_i \in V\}$ denota el conjunto de funciones asociadas al modelo M, las ecuaciones en $M_{\overline{a}}$ están dadas por $\{f_i^{\overline{a}} : v_i \in V\}$, definidas por

$$f_j^{\overline{a}}(x_{s_1}, x_{s_2,\ldots}, x_{s_j}, u_j) = f_j(x_{s_1}, x_{s_2,\ldots}, x_{s_j}, u_j) \quad \text{si } v_j \notin v_{i_1}, \ldots, v_{i_l} \tag{4.4.1}$$

$$f_j^{\overline{a}}(x_{s_1}, x_{s_2,\ldots}, x_{s_j}, u_j) = a_{i_s} \quad \text{si } v_j = v_{i_s}, \text{ para } 1 \leq s \leq l\ , \tag{4.4.2}$$

recordando que

$$pa_G(v_j) = \{v_{s_1} \ldots, v_{s_j}\}\ .$$

Notemos que las funciones en $M_{\overline{a}}$ resultan compatibles con el grafo $G_{\overline{A}}$, siendo $G_{\overline{A}}$ el grafo que se obtiene al eliminar en G todas las flechas que llegan a $\{v_{i_1}, \ldots, v_{i_l}\}$, los nodos asociados a las variables del vector A. Esquemáticamente, podemos representar a la intervención con

$$G_{\overline{A}} \qquad \Longleftrightarrow \qquad M_{\overline{a}} = \begin{cases} U = \{U_1, \ldots, U_n\} & \text{independientes,} \\ \{f_i^{\overline{a}} : v_i \in V\} & \text{definidas en (4.4.1) y (4.4.2)} \end{cases}$$

Notación: Para cada realización u del vector de errores U, notamos el valor tomado por las variables bajo el modelo M mediante $Y_M(u)$. Las variables construídas con el modelo $M_{\overline{a}}$ se denotarán indistintamente mediante $Y_{M_{\overline{a}}}$ o, para simplificar la notación, utilizaremos :

$$\begin{aligned} Y(u) &= Y_M(u) \\ Y_{\overline{a}}(u) &= Y_{M_{\overline{a}}}(u). \end{aligned} \tag{4.4.3}$$

$Y_{\overline{a}}$ representa el valor observado en cierta variables respuesta, tras haber intervenido. Es decir, $Y_{\overline{a}}$ es una variable contrafactual, ya que representan respuestas observadas en mundos intervenidos.

Ejemplo 4.4.1. *Consideremos un NPSEM para las variables (L, A, Y) donde, siguiendo la notación introducida en el Capítulo 2, L denota una covariable pre tratamiento, A representa el tratamiento asignado a cada individuo mientras que Y es la respuesta de interés observada. Si pensamos que el DAG asociado a estas variables está dado por la figura 4.4.1 el modelo M propone la siguiente representación para las variables observadas*

$$L = f_L(U_L) \tag{4.4.4}$$

$$A = f_A(L, U_A) \tag{4.4.5}$$

$$Y = f_Y(A, L, U_Y) \tag{4.4.6}$$

para ciertas funciones f_L, f_A, f_Y, y admite independencia entre las coordenadas del vector de perturbaciones $U = \{U_L, U_A, U_Z\}$. Para poder construir la respuesta contrafactual Y_a en el sentido de la

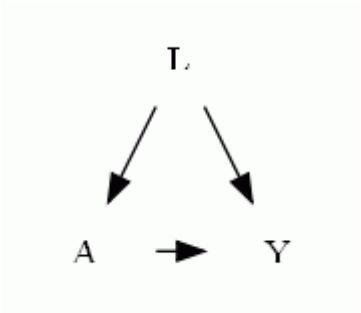

Figura 4.4.1: Grafo asociado al modelo M

Definición 2.3.2 introducida en el Capítulo 2, consideremos el modelo M_a, donde las variables aleatorias se definen iterativamente, con las mismas perturbaciones $U = \{U_L, U_A, U_Z\}$ utilizadas para definir las variables observadas, según las ecuaciones (4.4.4)-(4.4.6), pero utilizando ahora las funciones

$$f_L^a = f_L \tag{4.4.7}$$
$$f_A^a = a \tag{4.4.8}$$
$$f_Y^a = f_Y \tag{4.4.9}$$

compatibles con el grafo $G_{\overline{A}}$, dado por la figura 4.4.2. La variable asociada al nodo Y (ver la Sección 4.3

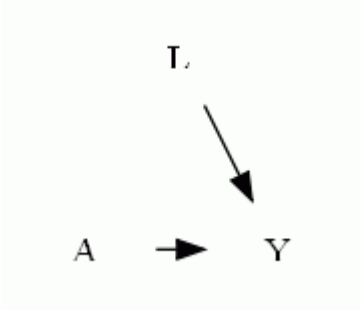

Figura 4.4.2: Grafo asociado al Modelo M_a

donde se discute el abuso notacional nodo-variable) construida a partir de las funciones (4.4.7)-(4.4.9) es la respuesta contrafactual Y_a.

4.5. Conexión entre contrafactuales y sem

A lo largo del Capítulo 2, en el marco de las respuestas contrafactuales, estudiamos condiciones bajo las cuales podíamos identificar el efecto medio del tratamiento. Los supuestos bajo los que identificamos son los siguientes: consistencia, positividad e intercambiabilidad. De estas tres condiciones, la única que se puede testear a partir de datos provenientes de un estudio observacional es la condición de positividad. Recordemos que, la misma establece que $0 < P(A = a|L = \ell) < 1$, cada vez que $P(L = \ell) > 0$. En el marco de las respuestas contrafactuales, las propiedades de intercambiabilidad y consistencia son llamadas "primitivas". En general, con este término haremos alusión a suposiciones que el investigador

está dispuesto a realizar en función del conocimiento específico que tiene del sistema en estudio, sabiendo que los datos no permitirán avalar ni refutar estas suposiciones.

En el marco de SEM, las "primitivas" son las ecuaciones estructurales y la función de probabilidad asignada a las perturbaciones. Es decir, el modelo M asume la existencia de funciones que dan origen a las variables observadas, junto con la independencia de las perturbaciones. Luego, se construyen las variables contrafactuales mediante los modelos intervenidos, como presentamos en la Definición 4.4.1. En este nuevo contexto, las condiciones de consistencia e intercambiabilidad condicional, o suposiciones semejantes que permitan garantizar la identificabilidad del parámetro causal de interés pueden ser deducidas a partir de las primitivas impuestas por el modelo M. Discutiremos estos aspectos siguiendo el trabajo de Pearl [20] , y por consiguiente, utilizando su notación, tal como advertimos en la Sección 4.3. Es decir, las letras mayúsculas denotarán indistintamente tanto variables aleatorias como nodos en el grafo.

Teorema 4.5.1. *Si M es un NPSEM asociado al DAG G, se tienen las siguientes propiedades:*

1. *Exclusión: Sea PA_Y el conjunto de padres de Y en el DAG G. Dado un conjunto W de nodos disjuntos con $\{Y\} \cup PA_Y$, tenemos que para todo valor pa_Y de PA_Y y w de W*

$$Y_{pa_Y,w} = Y_{pa_Y}$$

2. *Independencia: Si PA_J son los padres del nodo X_j en el DAG G y*

$$X_j^C = \left\{X_{j,pa_j} = f_j(pa_j, U_j) : \textit{con } pa_j \textit{ variando entre todos los posibles valores para } PA_J\right\}$$

entonces

$$X_1^C, X_2^C, ...X_k^C \textit{ son mutuamente independientes.}$$

La exclusión nos dice que, para cada nodo, el conjunto de padres incluye todas las variables que son causa directa de la variable asociada a dicho nodo. Por lo tanto, fijando el valor de las variables asociadas a los padres de un nodo, quedará determinado el valor de la variable asociada a dicho nodo, salvo por la perturbación correspondiente. Intervenir en cualquier otro nodo W no afectará al valor del nodo correspondiente a la variable Y.

La propiedad de independencia se hereda de la independencia entre las perturbaciones: entre las suposiciones hechas por el modelo M se inlcuye la independencia entre las coordenadas del vector $U = \{U_1, \ldots, U_n\}$, y por consiguiente tenemos la independencia entre las correspondientes variables contrafactuales. Esto se puede deducir observando que una vez que fijamos los valores de las variables en los nodos padres, la única fuente de aleatoriedad proviene del termino correspondiente a las perturbaciones:$X_{j,pa_j} = f_j(pa_j, U_j)$.

El próximo teorema resultará crucial a la hora de discutir la consistencia en este nuevo contexto:

Teorema 4.5.2. ***Composición:*** *Sea M un NPSEM asociado al DAG G. Dados X, Y, Z tres conjuntos disjuntos de nodos, sean x, z valores arbitrarios que pueden tomar las variables X y Z. Se tiene entonces la siguiente propiedad*

$$si\ Z_x = z \Longrightarrow Y_{x,z} = Y_x\ . \tag{4.5.1}$$

La composición afirma que en un mundo donde X se fija en x, si el valor tomado por Z_x, la variable asociada al nodo Z en este mundo, es z, entonces el valor de la variable asociada al nodo de Y en ese mundo (Y_x) sería el mismo valor que tomaría la correspondiente variable en un mundo en donde interviniésemos para fijar X en x y Z en z.

Lema 4.5.1. *Las variables construidas mediante un NPSEM y un NPSEM intervenido permiten verificar la condición de consistencia asumida en el contexto contrafactual, introducido en el Capítulo 2.*

Demostración: La composición es válida incluso tomando $X = \emptyset$. En tal caso tenemos que

$$Z_\emptyset = z \Longrightarrow Y_{\emptyset,z} = Y_\emptyset, \tag{4.5.2}$$

siendo $Z_\emptyset = Z$, $Y_\emptyset = Y$ mientras que $Y_{\emptyset,z} = Y_z$. Tenemos entonces que

$$\text{si } Z = z \Longrightarrow Y_z = Y, \tag{4.5.3}$$

tal como queríamos demostrar. □

Como hemos visto en el Capítulo 2, diferentes nociones de independencia o independencia condicional fueron necesarias para poder identificar parámetros causales. Hemos comentado al inicio de esta Sección que, en el contexto de NPSEM tales condiciones pueden ser deducidas, como muestra el siguiente lema.

Lema 4.5.2. *Consideremos el NPSEM con variables L, A, Y, introducido en el Ejemplo 4.4.1. Tenemos entonces que vale la aleatorización condicional:*

$$Y_a \coprod A \mid L \ .$$

Demostración: Para demostrar que se verifica la aleatorización condicional, usaremos la propiedad de *independencia* enunciada en el Teorema 4.5.1. En el presente contexto, la misma establece la independencia entre

$$L, A_\ell, Y_{a\ell} \ ,$$

para cada posible valor ℓ. Tenemos entonces que $Y_{a\ell} \coprod \{L, A_l\}$, de donde podemos deducir que $Y_{a\ell} \coprod A_\ell \mid L$, siendo que

$$\begin{aligned} P(Y_{a\ell} = y, A_\ell = b \mid L = \ell) &= \frac{P(Y_{a\ell} = y, A_\ell = b, L = \ell)}{P(L = \ell)} \text{ por independencia} \\ &= \frac{P(Y_{a\ell} = y)P(A_\ell = b, L)}{P(L = \ell)} \\ &= P(Y_{a\ell} = y)P(A_\ell = b \mid L = \ell) \text{ y como } Y_{al} \coprod \{L, A_\ell\} \Rightarrow Y_{a\ell} \coprod L \\ &= P\left(Y_{a\ell} = y \mid L = \ell\right) P\left(A_\ell = b \mid L = \ell\right) \ . \end{aligned}$$

Sea ℓ con $P(L = \ell) > 0$. Observemos que

1. a partir de los modelos M y M_a, tenemos que $L_a = L$,
2. por composición (ver ecuación (4.5.1)), $L_a = \ell$ implica que $Y_{a\ell} = Y_a$, es decir, en el presente contexto tenemos que en $L = \ell$, $Y_{\ell a} = Y_a$, siendo que $L_a = L$,
3. por consistencia (Lema 4.5.1), tenemos que en $L = \ell$ vale que $A_\ell = A$.

Podemos concluir entonces que para todo ℓ con $P(L = \ell) > 0$

$$Y_{a\ell} \coprod A_\ell \mid L = \ell \implies Y_a \coprod A \mid L = \ell \,,$$

tal como queríamos demostrar. □

El razonamiento efectuado a lo largo del Lema 4.5.2 admite importantes generalizaciones. En este sentido, cabe mencionar el criterio del *Back Door*, que presentaremos en la Sección 4.6. El mismo permite dar condiciones gráficas para garantizar la identificabilidad de la distribución de variables contrafactuales, asumiendo positividad.

Modelo contrafactual o NPSEM ?

A lo largo de esta Tesis hemos introducido dos manera posibles de abordar el problema causal. El modelo contrafactual, introducido en el Capítulo 2, y los modelos de ecuaciones estructurales no paramétricas (NPSEM). Cada uno de ellos presenta ventajas y desventajas, y no hay total consenso en la comunidad científica en favor de uno de ellos. El hecho de partir de diferentes supuestos repercute en el alcance de cada uno de los métodos. Además, NPSEM muchas independencias adicionales a las necesarias para identificar pueden ser deducidas a partir de la independencia de las perturbaciones.

4.6. Back Door

Como mencionamos en la Sección anterior, existe una herramienta gráfica que permite determinar en qué variables ajustar o condicionar para poder identificar la distribución de la variable contrafactual Y_a.

Teorema 4.6.1. *Back-Door (o de la puerta trasera) Sean A, L, Y tres conjuntos disjuntos de nodos en el grafo G. Consideremos la intervención en los nodos correspondientes A con el valor $\overline{a}$. Supongamos que se verifica la positividad y las siguientes condiciones:*

1. $an_G(L) \cap A = \emptyset$
2. *L bloquea todos los caminos de la puerta de atrás que van de A a Y, es decir,*

$$(A \amalg Y \mid L)_{G_{\underline{A}}} \,,$$

siendo $G_{\underline{A}}$ el grafo que obtenemos al eliminar de G todas las flechas que salen de los nodos en A

Tenemos entonces que

$$P(Y_{\overline{a}} = y) = \sum_\ell P(Y = y \mid L = \ell, A = \overline{a}).P(L = \ell). \tag{4.6.1}$$

La demostración se encuentra en el libro de J Pearl [20] (*Causality*, página 80). En la próxima Sección presentaremos una nueva demostración de este resultado para el caso en el que la intervención se produce en un único nodo.

La idea detrás de este teorema es que los caminos dirigidos a lo largo de las flechas entre A e Y transmiten relaciones causales entre A e Y, mientras que las rutas de acceso por la puerta trasera contienen las asociaciones entre las dos variables que hacen que la medida de asociación no sea igual a la medida causal. Por lo tanto, el bloqueo de tales caminos, asegura que el efecto que se tiene a partir de la distribución de las variables observadas hace al parámetro causal identificable.

El teorema de back door responde tres cuestiones:

1. ¿Existe confusión? La respuesta es afirmativa si existen caminos por la puerta de atrás entre el tratamiento y la respuesta que no conseguimos bloquear con las variables medidas.

2. ¿Puede ser la confusión eliminada? Esto ocurre si todos los caminos por la puerta de atrás entre el tratamiento y la respuesta pueden ser bloqueados usando variables medidas.

3. ¿Qué variables son necesarias para eliminar la confusión? El Teorema 4.6.1 permite decidir qué variables necesitamos medir para que todos los caminos por la puerta de atrás entre el tratamiento y la respuesta estén bloqueados por tales variables y podamos así identificar la distribución de la variable contrafactual.

Ejemplo 4.6.1. *El teorema de ajuste back door*

Consideremos el grafo de la figura 4.6.1 que fue utilizado para ilustrar los modelos gráficos y sus aplicaciones en epidemiología por Greenland et al en [7]

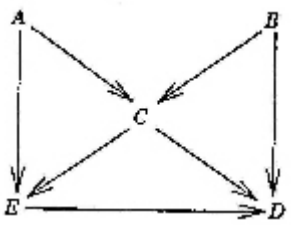

Figura 4.6.1: A= indicador de polución del aire, B=sexo, C=actividad bronquial, E=tratamiento antihistamínico, D=asma.

Examinemos que variables satisfacen el criterio back door para el par (E, D), interviniendo en E.
A no satisface el criterio pues no bloquea el camino E, C, D
B no satisface el criterio por la misma razón
C no lo satisface porque desbloquea el camino E, A, C, B, D
(A, C) sí lo satisface y también el par (B, C)
Entonces podemos conlcuir que:

$$\begin{aligned} P(D_e = d) &= \sum_a \sum_c P(D = d \mid E = e, A = a, C = c).P(A = a, C = c) \\ &= \sum_b \sum_c P(D = d \mid E = e, B = b, C = c).P(B = b, C = c) \end{aligned}$$

Luego para identificar a $P(D_e = d)$,además de medir E y D, es suficiente observar A y C o bien E y C, pues ambos pares bloquean todos los caminos por la puerta trasera.

4.6.1. Intervención alternativa

Procuraremos en esta Sección hacer una nueva propuesta para representar intervenciones, de forma que con el nuevo sistema de ecuaciones podamos construir de manera simultánea a la variable tratamiento y a la variable contrafactual, al menos cuando la intervención se produce en un único nodo. Con esta nueva construcción podremos demostrar que las condiciones del Teorema 4.6.1 garantizan las condiciones de identificabilidad requeridas en el Capítulo 2, en lo que a aleatorización condicional respecta.

Sea $A = A_1$ la variable en la que queremos intervenir, asociada al nodo v_{i_1}. Dado un NPSEM M, caracterizado por U y $\{f_j : v_j \in V\}$, el nuevo modelo $M_{\text{NEW}a}$ de intervención está conformado por la mismas perturbaciones U que en el modelo original, mientras que en lugar de hacer constante la función correspondiente al nodo v_{i_1}, propagaremos el efecto de la intervención a través de sus hijos, poniendo como entrada en los hijos del nodo v_{i_1} el valor a de la constante con la que deseamos intervenir. Mas específicamente, sea $G_{\underline{A}}$ el grafo en el que eliminamos todas las flechas que salen del vértice $A_1 = v_{i_1}$. El sistema de funciones asociado a la nueva intervención resultará compatible con el grafo $G_{\underline{A}}$. Ahora, al actualizar el valor de una variable asociada al nodo v_j que requiera del valor de las variables correspondiente al nodo v_{i_1}, es decir, si $v_{i_1} \in pa_G(v_j)$, utilizaremos siempre el valor a. Tenemos entonces la siguiente manera alternativa para representar una intervención.

Definición 4.6.1. Sea M un modelo causal de ecuaciones estructurales no paramétricas para las variables $X = (X_1, \ldots, X_n)$, con funciones estructurales $\{f_i : v_i \in V\}$ y perturbaciones $U = (U_1, U_2, ...U_n)$.

Dada la variable A asociada al nodo v_{i_1} y un posible valor a para la variable A, definimos $M_{\text{NEW}a}$ siendo el NUEVO modelo intervenido, dado por:

1. el mismo vector $U = \{U_1, \ldots, U_n\}$ de perturbaciones que en el modelo M (Definición 4.2.1)
2. un nuevo conjunto de ecuaciones, que coincide con las del modelo M excepto en lo que respecta a los nodos correspondientes a los hijos de la variable en la que queremos intervenir. Es decir, las funciones $\{f_j^{\text{NEW}a} : v_j \in V\}$ están dadas por

   a) el mismo vector $U = \{U_1, \ldots, U_n\}$ de perturbaciones que en el modelo M (Definición 4.2.1)

   b) un nuevo conjunto de ecuaciones, que coincide con las del modelo M excepto en lo que respecta a los nodos correspondientes a los hijos de la variable en la que queremos intervenir. Es decir, las funciones $\{f_j^{\text{NEW}a} : v_j \in V\}$ están dadas por

$$f_j^{\text{NEW}a} = f_j \quad si\ v_{i_1} \notin pa_G(v_j) \tag{4.6.2}$$
$$f_j^{\text{NEW}a} = f_j(a, x_{s_2}..., x_{s_j}, u_j) \quad si\ v_{i_1} = v_{s_1} \in pa_G(v_j)\ . \tag{4.6.3}$$

   siendo

$$pa_G(v_j) = \{v_{s_1}, \ldots, v_{s_j}\}.$$

   Notemos que las funciones en $M_{\text{NEW}a}$ resultan compatibles con el grafo $G_{\underline{A}}$, siendo $G_{\underline{A}}$el grafo que se obtiene al eliminar en G todas las flechas que salen de v_{i_1}, el nodo asociados a la variable A. Esquemáticamente, podemos representar a la NUEVA intervención con

$$G_{\underline{A}} \qquad \Longleftrightarrow \qquad M_{\text{NEW}a} = \begin{cases} U = \{U_1, \ldots, U_n\} & \text{independientes,} \\ \{f_j^{\text{NEW}a} : v_j \in V\} & \text{definidas en (4.6.2) y (4.6.3).} \end{cases}$$

Notación: Utilizaremos Y^a para denotar a las variables construídas utilizando el modelo $M_{\text{NEW}a}$. Es decir, tenemos

$$\begin{aligned} Y(u) &= Y_M(u) \\ Y_a(u) &= Y_{M_a}(u) \\ Y^a(u) &= Y_{M_{\text{NEW}a}}(u) \end{aligned}$$

Dando continuidad al Ejemplo 4.4.1, consideremos el grafo G con nodos (L, A, Y). Las variables asociadas al modelo M, M_a y $M_{\text{NEW}a}$, compatibles con G, $G_{\overline{A}}$ y $G_{\underline{A}}$, respectivamente, están dadas por los siguientes sistemas:

$$L = f_L(U_L) \qquad L_a = f_L(U_L) \qquad L^a = f_L(U_L) \tag{4.6.4}$$

$$A = f_A(L, U_A) \qquad A_a = a \qquad A^a = f_A(L^a, U_A) \tag{4.6.5}$$

$$Y = f_Y(A, L, U_Y) \quad Y_a = f_Y(A_a, L_a; U_Y) \quad Y^a = f_Y(a, L^a, U_Y) \tag{4.6.6}$$

Observar los grafos de las figuras 4.6.2, 4.6.3 y 4.6.4.

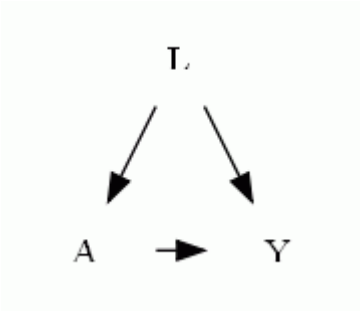

Figura 4.6.2: Grafo G asociado al modelo M

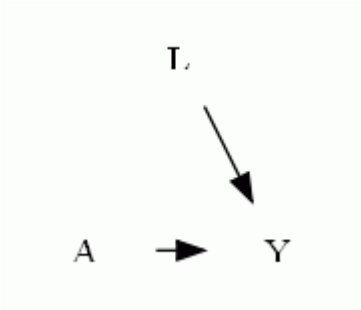

Figura 4.6.3: Grafo $G_{\overline{A}}$asociado al modelo intervenido segun Pearl M_a

La ventaja esperada con la nueva forma de representar la intervención radica en la capacidad de representar a la variable tratamiento A, a la variable contrafactual Y_a y a L mediante un mismo sistema de ecuaciones, como se prueba en el siguiente resultado.

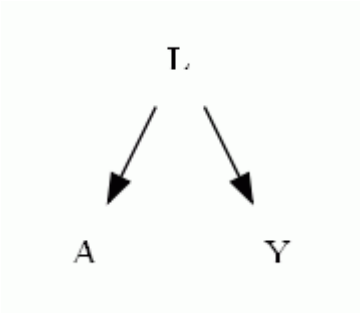

Figura 4.6.4: Grafo $G_{\underline{A}}$ asociado al modelo intervenido M_{NEWa}

Lema 4.6.1. *Consideremos las variables construídas bajo M, M_a y M_{NEWa}, introducidas en el Ejemplo 4.4.1. Tenemos entonces que $Y_a \coprod A|L$.*

Demostración: A partir de la construcción de los vectores (L, A, Y), (L_a, A_a, Y_a) y (L^a, A^a, Y^a), presentadas en (4.6.4)-(4.6.6), tenemos que

1. $L = L^a = L_a$,
2. $Y_a = Y^a$,
3. $A^a = A$.

Por otra parte, tenemos que L d-separa A de Y en $G_{\underline{A}}$ y por consiguiente, $Y^a \coprod A^a|L^a$.
De las observaciones 1-3, concluimos que

$$Y_a \coprod A|L \ .$$

□

El resultado anterior puede ser generalizado sin mayor dificultad, en la medida que $an_G(L) \cap A = \emptyset$ y L d-separe A de Y en $G_{\underline{A}}$. Estas son las condiciones bajo las cuales el Teorema 4.6.1 demuestra que vale la identificabilidad. Siguiendo el abordaje presentado en el Capítulo 2, habiendo probado que las primitivas asociadas a los NPSEM permiten deducir la condición de consistencia, si vale la positividad, resta verificar la aleatorización condicional para identificar la distribución de las respuestas contrafactuales. El siguiente resultado prueba que bajo las condiciones del Teorema 4.6.1 vale la aleatorización condicional.

Teorema 4.6.2. *Sea M un NPSEM que deseamos intervenir en un único nodo A con el valor a. Consideremos las variables construídas bajo M, M_a y M_{NEWa}. Bajo las condiciones del Teorema 4.6.1, tenemos que*

$$Y_a \coprod A|L \ .$$

Si además se verifica la condición de positividad también obtenemos la fórmula (4.6.1) para identificar la distribución contrafactual.

Demostración: Los siguientes hechos se deducen de la construcción de M_a y M^a a partir de M:

1. Si $an_G(L) \cap A = \emptyset$, entonces $L = L^a = L_a$.
2. $A = A^a$. Es en este punto donde utilizamos que la intervención se produce en un único nodo.

3. $Y_a = Y^a$.

Luego, para verificar la independencia condicional, basta ver que Y^a es independiente de A^a dado L^a. Estas variables han sido construídas bajo el modelo M^a y, por construcción, su distribución es compatible con la del grafo $G_{\underline{A}}$. Como en este grafo tenemos que L d-separa A de Y, vale que

$$Y^a \coprod A^a | L^a \ ,$$

y por consiguiente, $Y^a \coprod A|L$, como se quería demostrar. La fórmula 4.6.1 se deduce del Lema 2.5.1.

□

Bibliografía

[1] Cole, S. & Frangakis, C. (2009).The Consistency Statement in Causal Inference: A definition or an Assumption?. *Epidemiology,* **20 (1),** 3-5

[2] Cox, D. R. & Wermuth, N. (2004). Causality: a Statistical View. *International Statistical Review,* **72 (3),** 285-305

[3] Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal Royal Statistical Society. Series B (methodological)*, **41 (1),** 1-31

[4] Dawid, A. P. (2007). *Fundamentals of Statistical Causality. Research Report* **279**. Department of Statistical Science, University College London. 94 pp.

[5] Ferrari, P.A. y Galves, A. (1997). *Acoplamiento e procesos estocásticos*. Rio de Janeiro: SBM, IMPA.

[6] Geiger D. , Verma, Y. & Pearl, J. (1990). Identifying Independence in Bayesian Networks.UCLA Cognitive System Laboratory, Technical Report CSD-890028. *Networks***, 20 (5),** 507-534

[7] Greenland, S. , Pearl, J. And Robins J. (1999). Causal diagrams for epidemiologic research. *Epidemiology,* **10**, 37-48

[8] Heckman J. J. and Hotz V. J. (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association,* **84**, 862–74.

[9] Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health,* **58,** 265-271

[10] Hernán, M. A. , Hernández-Díaz S. & Robins J. M. (2004). A structural approach to selection bias. *Epidemiology***, 15** (5), 615-625.

[11] Hernán, M. A. and Robins J.M. (2006). Estimating causal effects from epidemiological data. *Journal Epidemiology and Community Health,* **60**, 578-586.

[12] Hernán, M. A. & Robins, J. M. (Aparecerá en 2011). *Causal Inference*. London: Chapman & Hall/CRC.

[13] Hernández-Díaz, S., Schisterman, E. F. and Hernán,M. A. (2006). The Birth Weight "Paradox" Uncovered?. *American Journal of Epidemiology,* **164,** 1115-1120

[14] Holland, P.W. (1986).Statistics and Causal Inference.*Journal of the American statistical Association,* **81**, N°396, 945-960.

[15] Lauritzen, S.L. (1996). *Graphical models*. Oxford,UK: Oxford University Press, Clarendon.

[16] Little, R. & Rubin, D. (2000). Causal Effect in Clinical and Epidemiological Studies Via Potential Outcome:Concepts and Analytical Approaches.*Annual Review of Public Health,* **21**, 121-145.

[17] Morgan, S (2001). Conterfactuals, Causal Effect, Heterogeineity and the Catholic School Effect on Learning. *Sociology of Education,* **74**, 341-374.

[18] Morgan, S & Winship, C.(2007). *Counterfactuals and Causal Inference.Methods and Principles for Social Research,* New York: Cambridge University Press.

[19] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*.San Francisco, CA: Morgan Kaufmann.

[20] Pearl, J. (2000). *Causality: models, reasoning and infererence*. New York : Cambridge University Press.

[21] Pearl, J (2010). An Introduction to Causal Inference. *The International Journal of Biostatistics*, **6**, Iss. 2, Article 7.

[22] Rotnitzky, A. (2009).Notas del curso "Inferencia causal" en el X Congreso Monteiro disponibles en http://www.matematica.uns.edu.ar/XCongresoMonteiro/Docs/inferencia-causal-andrea.pdf

[23] Rubin, D.(1974). Estimating Causal Effect of Treatments in Randomized and Non randomized Studies. *Journal of Educational Psychology*, **66** (5), 688-701.

[24] Rubin,D. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association,* **75**, No. 371, 591-593.

[25] Verma, T. & Pearl, J. (1988) Causal network: semantics and expressiveness. *In Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence,* (Mountain View, CA), pp. 352-9. Reprinted in R.Shachter, T. S. Levitt, and L. N. Kanal (Eds.)(1990), *Uncertainty in Artificial Intelligence*, **4**, 69-76 Amsterdam: Elesevier.

[26] Wasserman, L.(2004). *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer Text in Statistics.

[27] Wilcox, A. J. (2006) Invited Commentary: The Perils of Birth Weight—A Lesson from Directed Acyclic Graph.*American Journal of Epidemiology,* **164,** 1121-1123.