

# Fundamentos Causalidad I

Alvaro J. Riascos Villegas

Marzo de 2020

# Contenido

- 1 Introducción
- 2 Paradoja de Simpson
- 3 Probabilidad y Estadística Básica
- 4 Grafos
- 5 Modelos Causales Estructurales
- 6 Modelos de Causalidad

- Muchas preguntas que frecuentemente nos hacemos de un conjunto de datos: cómo y por qué, no es posible responderlas en el marco estadístico tradicional.
- Es necesario un marco conceptual adicional, una teoría de causalidad.

# Contenido

- 1 Introducción
- 2 Paradoja de Simpson**
- 3 Probabilidad y Estadística Básica
- 4 Grafos
- 5 Modelos Causales Estructurales
- 6 Modelos de Causalidad

- Edward Simpson (1922): Una relación estadística que se cumple para toda la población puede reversarse en cada subpoblación.
- Por ejemplo: en una población de estudiantes se podría encontrar que en promedio los fumadores tienen mejores notas. Sin embargo, entre cada grupo de edad los fumadores tienen peores notas y, entre cada grupo de edad y sexo los fumadores tiene mejores notas y así puede reversarse la asociación anterior en cada subpoblación.

## Example

Se les ofrece tomar de forma voluntaria una droga a 700 pacientes. 350 pacientes la toman y los demás no.

**Table 1.1** Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

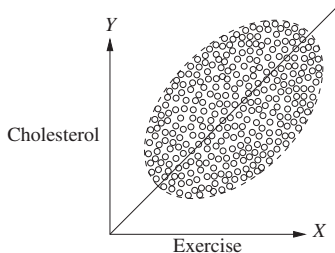
Los datos sugieren que si conocemos el sexo de las personas, debemos recomendar tomar la droga. Pero si no lo conocemos, no!

# Paradoja de Simpson: Ejemplo

- En esta ejemplo tenemos 343 mujeres y 357 hombres.
- Es imposible racionalizar este fenómeno sin apelar a alguna teoría (i.e., hipótesis):
  - Suponga que el estrógeno reduce la efectividad de la droga. Sin embargo, supongamos que esta es efectiva tanto en hombres como mujeres.
  - Las mujeres son las que mas optan por tomar la droga.
  - El efecto poblacional de tomar la droga se puede obtener de repetir y promediar muchas veces este ejercicio aleatoriamente: elegir aleatoriamente una persona que tomo la droga y comparala contra una persona elegida aleatoriamente que no tomo la droga.
  - Como es más probable elegir una mujer en el primer caso y un hombre en el segundo caso, parece a nivel poblacional que la droga no funciona.

# Paradoja de Simpson: Ejemplo

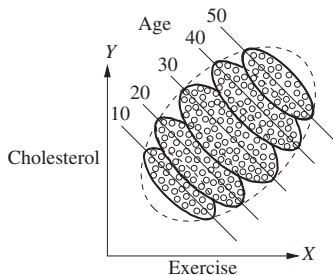
- Una versión continua de esta paradoja es la siguiente. Considere la relación entre el colesterol y el ejercicio de un grupo de personas.



**Figure 1.2** Results of the exercise–cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

# Paradoja de Simpson: Ejemplo

- Ahora si condicionamos a la edad la historia es muy distinta.



**Figure 1.1** Results of the exercise–cholesterol study, segregated by age

# Paradoja de Simpson: Ejemplo

- De nuevo, la única forma de racionalizar esto es apelando a alguna teoría (historia) que va más allá de los datos disponibles.
- Suponga que los mayores hacen más ejercicio que los más jóvenes y tiene el colesterol más alto independientemente de si hacen ejercicio.
- Entonces la edad es un factor común del ejercicio y los niveles de colesterol.

# Paradoja de Simpson: Desagregar los datos no es siempre la solución

- Suponga que en el primer ejemplo no se registra el sexo. Se registra la presión arterial después del experimento.
- Suponga que la droga afecta la recuperación reduciendo la presión pero también tiene un efecto tóxico. En cada subgrupo el efecto tóxico predomina.
- Los resultados del experimentos son como a continuación (esta gráfica es idéntica a la anteriores excepto que se intercambiaron los nombres de las columnas).

# Paradoja de Simpson: Ejemplo

- Los resultados a nivel agregado sugieren que se debe usar la droga, pero no si se analizan los datos condicional a la presión arterial.
- Si la presión arterial se midiera antes del experimento y esta tuviera un efecto sobre las personas que acceden al tratamiento y no al contrario, encontraríamos lo contrario,

**Table 1.2** Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

# Paradoja de Simpson: Conclusiones

- En conclusion: En algunos casos la respuesta correcta puede estar en los datos desagregados, como en el primer ejemplo y en otros casos en la informacion agregada como en el segundo ejemplo.
- En la paradoja de Simpson, el problema está en asumir que el tratamiento no tiene ningún efecto sobre el sexo de las personas que acceden voluntariamente (i.e., que no hay autoselección).
- El análisis estadístico solo no permite responder estas preguntas, es necesario hipótesis adicionales y una teoría de como se generan los datos.

# Correlación y Causalidad

- Crimen y consumo de helados, ocupación hotelera y precios.
- Muchos ejemplos de correlación: incendios y número de bomberos, personas que andan apuradas y llegada tarde a reuniones, etc.

# Contenido

- 1 Introducción
- 2 Paradoja de Simpson
- 3 Probabilidad y Estadística Básica**
- 4 Grafos
- 5 Modelos Causales Estructurales
- 6 Modelos de Causalidad

# Probabilidad Condicional y Ley de Bayes

- La paradoja del gato (Monty Hall) tiene origen en que en ocasiones nuestra creencias solo dependen de los datos observados y desconocen la forma como se generan esos datos.
- Cuando el maestro de ceremonias abre la puerta donde no está el gato esto no es por si solo informativo hasta tanto no se cuestione el procesos de decisión que siguió este para abrir la puerta.

# Observaciones sobre el método de regresión lineal

- Si especificamos la relación entre dos variables como:

$$y = \beta x + \epsilon \quad (1)$$

entonces el estimador de mínimos cuadrados ordinarios es:

$$\beta = \frac{\sigma_{xy}}{\sigma_x^2}$$

# Contenido

- 1 Introducción
- 2 Paradoja de Simpson
- 3 Probabilidad y Estadística Básica
- 4 Grafos**
- 5 Modelos Causales Estructurales
- 6 Modelos de Causalidad

- Pariente, hijo, descendientes.

# Contenido

- 1 Introducción
- 2 Paradoja de Simpson
- 3 Probabilidad y Estadística Básica
- 4 Grafos
- 5 Modelos Causales Estructurales**
- 6 Modelos de Causalidad

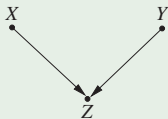
- Un modelo causal consiste de un conjunto de variables aleatorias  $U$  y  $V$  y un conjunto de funciones  $F$  que establecen una relación entre las variables de  $U$  y  $V$ .

## Example (Salario, educación y experiencia)

Sea  $U = \{X, Y\}$ ,  $V = \{Z\}$  y  $F = \{f_Z\}$ , tal que  $f_Z$  es la distribución de  $Z = 2X + 3Y$ . El modelo representa la relación de causalidad entre  $X$ , años de escolaridad,  $Y$  experiencia profesional y  $Z$  salario.

- Todo SCM define se puede representar como un modelo gráfico (grafos dirigidos acíclicos).

## Example (Salario, educacion y experiencia)



**Figure 1.9** The graphical model of SCM 1.5.1, with  $X$  indicating years of schooling,  $Y$  indicating years of employment, and  $Z$  indicating salary

- Cuando un grafo es acíclico sobre variables aleatorias  $\{X_1, \dots, X_n\}$  la distribución de probabilidad conjunta de  $(X_1, \dots, X_n)$  se puede descomponer como:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i \mid pa_i)$$

donde  $pa_i$  son los parientes de  $X_i$ .

# Contenido

- 1 Introducción
- 2 Paradoja de Simpson
- 3 Probabilidad y Estadística Básica
- 4 Grafos
- 5 Modelos Causales Estructurales
- 6 Modelos de Causalidad**

Example (Financiación de la Escuela, SAT Promedio, Tasa Aceptación Universidad)

Sea  $V = \{X, Y, Z\}, U = U_x, U_y, U_z$  y  $F = \{f_x, f_y, f_z\}$ :

$$x = U_x$$

$$y = \frac{x}{3} + U_y$$

$$z = \frac{y}{16} + U_z$$

Example (Horas de Trabajo, Entrenamiento, Tiempo de Carreras)

Sea  $V = \{X, Y, Z\}, U = U_x, U_y, U_z$  y  $F = \{f_x, f_y, f_z\}$  y:

$$x = U_x$$

$$y = 84 - x + U_y$$

$$z = \frac{100}{y} + U_z$$

- Los dos ejemplos anteriores se pueden representar gráficamente como:

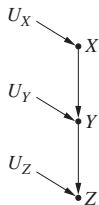


Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

- Este es un ejemplo de una **cadena**. Decimos que  $Z$  es potencialmente dependiente de  $Y$ ,  $Y$  es potencialmente dependiente de  $X$ ,  $Z$  es potencialmente dependiente de  $X$  y  $Z$  es independiente de  $X$  condicional en  $Y$ .

## Definition

Dos variables  $X$  y  $Z$  son condicionalmente independientes dado  $Y$  si existe solo un camino unidireccional entre  $X$  y  $Z$ , y  $Y$  es cualquier conjunto de variables que intercepta este camino.

## Example (Temperatura, Helados y Crimen)

Sea  $V = \{X, Y, Z\}$ ,  $U = U_x, U_y, U_z$  y  $F = \{f_x, f_y, f_z\}$  y:

$$x = U_x$$

$$y = 4x + U_y$$

$$z = \frac{x}{10} + U_z$$

- El ejemplo anterior se pueden representar gráficamente como:

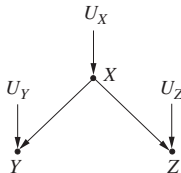


Figure 2.2 The graphical model of SCMs 2.2.5 and 2.2.6

- Este es un ejemplo de una **bifurcación**.
- Decimos que  $Z$  es potencialmente dependiente de  $X$ ,  $Y$  es potencialmente dependiente de  $X$ ,  $Z$  es potencialmente dependiente de  $Y$ ,  $Z$  es independiente de  $Y$  condicional en  $X$ .

# Modelos Gráficos: Independencia Condicional en Bifurcaciones

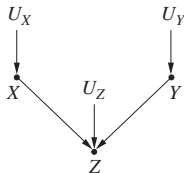
## Definition

Si una variable  $X$  es una casusa común de  $Y$  y  $Z$  y solo hay un camino entre  $Y$  y  $Z$  entonces  $Y$  y  $Z$  son independientes condicional a  $X$ .

# Modelos Gráficos Asociados: Colisionadores

- Sea  $Z = X + Y$  donde  $X$  y  $Y$  son independientes.
- Por ejemplo  $X, Y$  pueden ser el resultado de lanzar aleatoriamente una moneda al aire. Si por lo menos una cae cara  $Z$  se activa, suena una campana.
- Sea  $X$  la habilidad musical de un individuo y  $Y$  su desempeño académico.  $X, Y$  son posiblemente independientes. Ahora sea  $Z$  si la persona tiene una beca de estudio o no.

- Lo ejemplos anteriores se pueden representar gráficamente como:



**Figure 2.3** A simple collider

- $X$  y  $Z$ , y  $Y$  y  $Z$  son dependientes.  $X$  y  $Y$  son independientes y  $X$  y  $Y$  dado  $Z$  son dependientes.

# Modelos Gráficos: Independencia Condicional en Colisionadores

## Definition

Si una variable  $Z$  es una variable de colisión entre dos variables  $X$  y  $Y$ , y existe solo un camino entre  $X$  y  $Y$ , entonces  $X$  y  $Y$  son incondicionalmente independientes, pero son dependientes condicionalmente en  $Z$  y cualquiera de sus descendientes.

# Modelos Gráficos: Independencia Condicional en Colisionadores

- En el ejemplo de la paradoja del gato (i.e., Monty Hall problem) el nodo  $X$  es la puerta que el agente elige, el nodo  $Y$  es donde esta el gato y  $Z$  es la puerta que abre el maestro de ceremonia.
- Si  $X$  es distinto a  $Y$ ,  $Z$  queda completamente determinado.

# Modelos Gráficos: Independencia Condicional en Colisionadores

## Example ( $Z = X+Y$ )

$X, Y$  Bernoulli independientes.

**Table 2.2** Conditional probability distributions for the distribution in Table 2.2. (Top: Distribution conditional on  $Z = 1$ . Bottom: Distribution conditional on  $Z = 0$ )

$X$	$Y$	$P(X, Y Z = 1)$
Heads	Heads	0.333
Heads	Tails	0.333
Tails	Heads	0.333
Tails	Tails	0

$X$	$Y$	$Pr(X, Y Z = 0)$
Heads	Heads	0
Heads	Tails	0
Tails	Heads	0
Tails	Tails	1

# Modelos Gráficos: Independencia Condicional en Colisionadores

- La condición de dependencia condicional a un colisionador es importante para determinar si un modelo causal generó un conjunto de datos, para descubrir modelo a partir de los datos y para resolver la Paradoja de Simpson.
- Esta propiedad deja claro una nueva forma en la que dos variables pueden ser dependientes:
  - 1 Una variable causa la otra.
  - 2 Una tercera variable causa ambas.
  - 3 Dependencia condicional a  $Z$ . Obérvase que en esta última  $X$  y  $Y$  son independientes.

# Modelos Gráficos: d-separación (separación direccional)

- El concepto de d-separación sirve como criterio para explorar las dependencias en cualquier modelo gráfico.
- Dos nodos que están d-separados son independientes. Si no están d-separados entonces son potencialmente dependientes.

## Definition (d-separación)

Decimos que un camino  $p$  es bloqueado por un conjunto de nodos  $Z$  si se cumple alguna de estas:

- 1  $p$  contiene un cadena  $A \rightarrow B \rightarrow C$  donde  $B \in Z$ .
- 2  $p$  contiene un bifurcación  $A \leftarrow B \rightarrow C$  donde  $B \in Z$ .
- 3  $p$  contiene un colisionador  $A \rightarrow B \leftarrow C$  donde  $B \notin Z$  y ningún descendiente de  $B$  está en  $Z$ .

Si  $Z$  bloquea todos los caminos de  $X$  a  $Y$  entonces  $X$  y  $Y$  están d-separados condicional en  $Z$  y por lo tanto  $X$  y  $Y$  son independientes condicional en  $Z$ .

## Example (Z,Y d-separados condicional al conjunto vacío)

Solo hay un camino entre  $Z$  y  $Y$  y ese camino lo bloquea un colisionador:  $Z \rightarrow W \leftarrow X$ . Luego  $Z$  y  $Y$  son incondicionalmente independientes (i.e., independientes condicional al conjunto de nodos vacío).

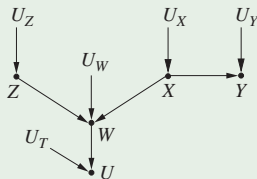


Figure 2.7 A graphical model containing a collider with child and a fork

## Example (Z,Y d-conectados condicionales a W)

Solo hay un camino entre Z y Y. Si se condiciona a W queda una bifurcación X que no está en el conjunto W y el único colisionador si está luego ese camino no está bloqueado.

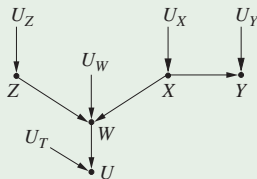
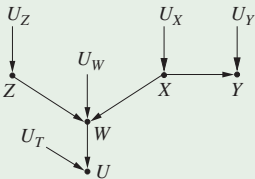


Figure 2.7 A graphical model containing a collider with child and a fork

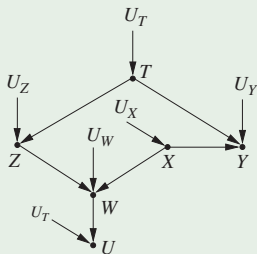
Example ( $Z, Y$  d-separados condicional a  $\{W, X\}$ )

$\{W, X\}$  bloquea e único camino entre  $Z$  y  $Y$ .



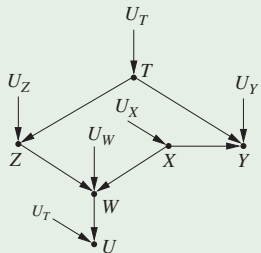
**Figure 2.7** A graphical model containing a collider with child and a fork

## Example (Z,Y incondicionalmente dependientes)



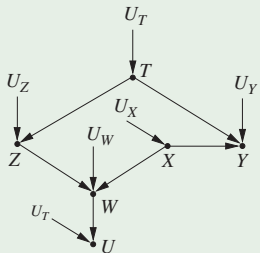
**Figure 2.8** The model from Figure 2.7 with an additional forked path between Z and Y

## Example (Z,Y condicional a $T$ independientes)



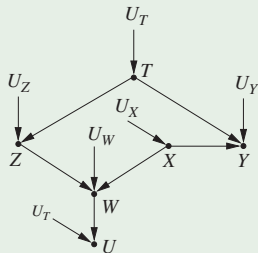
**Figure 2.8** The model from Figure 2.7 with an additional forked path between  $Z$  and  $Y$

## Example (Z,Y condicional a T, W d-conectados)



**Figure 2.8** The model from Figure 2.7 with an additional forked path between  $Z$  and  $Y$

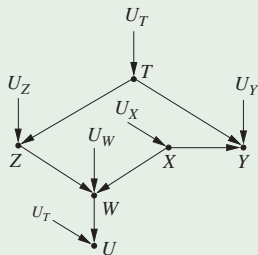
Example (Z,Y condicional a T, W, X independientes)



**Figure 2.8** The model from Figure 2.7 with an additional forked path between Z and Y

## Example (Z,Y Resumen de Dependencias)

$Z$ ,  $Y$  son potencialmente dependientes condicional a  $W, U, \{W, U\}, \{W, T\}, \{U, T\}, \{W, U, T\}, \{W, X\}, \{U, X\}$  y  $\{W, U, X\}$ . Son d-separados o condicionalmente independientes en  $T, \{X, T\}, \{W, X, T\}, \{U, X, T\}$  y  $\{W, U, X, T\}$ .



**Figure 2.8** The model from Figure 2.7 with an additional forked path between  $Z$  and  $Y$

# Pruebas de Especificación de Modelos e Identificación Causalidad

- Supongamos que conjeturamos que el modelo probabilístico que generó los datos es:

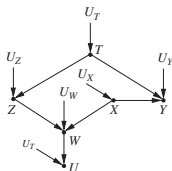


Figure 2.8 The model from Figure 2.7 with an additional forked path between Z and Y

- $W, Z$  son independientes condicionales a  $Z_1$ . Esto se puede probar mediante una regresión:  $w = r_X x + r_1 z_1$ . Si  $r_1$  es estadísticamente diferente a cero, entonces la hipótesis se rechaza, y el modelo gráfico no lo validan los datos (prueba local).