PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos y técnicas de investigación cuantitativa

César Antonio Aguilar

Facultad de Lenguas y Letras

29/04/2013

Cesar.Aguilar72@gmail.com

Revisión de tarea (1)

Para iniciar, vamos a resolver la tarea: analizar las frecuencias de distribución para las secuencias de palabras (o términos) *base de datos/base de conocimiento*.

Antes de hacer este cálculo, les propongo una fórmula para representar probabilidades condicionales. Veamos primero **base de datos**:

$$P(Dato \mid Base\_de) = P(Dato \cap Base\_de) \,/\, P(Base\_de)$$

Lo cual se lee como: *la probabilidad de que ocurra **dato** ligado con la secuencia **base_de**, es igual a la probabilidad de que ocurra **dato** con **base_de**, entre la probabilidad de que ocurra **base_de**.*

Revisión de tarea (2)

Ahora, hagamos lo mismo para **base de conocimiento**:

$$P(\text{Conocimiento} \mid \text{Base\_de}) = P(\text{Dato} \cap \text{Base\_de}) / P(\text{Base\_de})$$

Lo cual se lee como: *la probabilidad de que ocurra **conocimiento** ligado con la secuencia **base_de**, es igual a la probabilidad de que ocurra **conocimiento** con **base_de**, entre la probabilidad de que ocurra **Base**.*

Revisión de tarea (4)

Una vez hecho lo anterior, pasemos a hacer nuestra exploración de casos con **Bwananet**. Para ello, les propongo empezar del siguiente modo:

Revisión de tarea (5)

En esta primera búsqueda en el área de *Informática*, lo que hacemos es ubicar con qué nombre se asocia frecuentemente la secuencia **base_de**. Así, tenemos lo siguiente:

Secuencias generadas	Total de casos
Base_de **datos**	43
Base_de **conocimiento**	5
Base_de **información**	1
Base_de **periódicos**	1

Revisión de tarea (6)

En contraste, para el área de *Genómica* tenemos:

Secuencias generadas	Total de casos
Base_de **datos**	37
Base_de **ADN**	3
Base_de **longitud**	2
Base_de **experimentos**	1
Base_de **molde**	1
Base_de **heterogeneidad**	1
Base_de **nucleótidos**	1
Base_de **guanina**	1
Base_de **secuenciación**	1
Base_de **purina**	1
Base_de **huesos**	1

Revisión de tarea (7)

Si contrastamos ahora los datos que generan las secuencias **base_de datos** *versus* **base_de conocimientos** tenemos:

Área	Términos	Total
Informática	Base_de **datos**	50
	Base_de **conocimiento**	50

Área	Términos	Total
Genómica	Base_de **datos**	50
	Base_de **conocimiento**	0

Revisión de tarea (8)

Los adjetivos que se combinan con **base_de_datos** en *informática* son:

Secuencias generadas	Total de casos
Base_de datos **documentales**	7
Base_de datos **relacionales**	6
Base_de datos **bibliográfica**	4
Base_de datos **jurídica**	1
Base_de datos **local**	3
Base_de datos **remota**	2
Base_de datos **internacionales**	1
Base_de datos **global**	2
Base_de datos **corporativas**	1
Base_de datos **original**	1
Base_de datos **temporal**	1
Base_de datos **multiusuario**	1
Base_de datos **subyacente**	1

Revisión de tarea (9)

Secuencias generadas	Total de casos
Base_de datos **agrarias**	1
Base_de datos **estadísticos**	1
Base_de datos **operacional**	1
Base_de datos **actual**	1
Base_de datos **jerárquica**	3
Base_de datos **compatible**	1
Base_de datos **accesible**	1
Base_de datos **objeto-relacional**	1
Base_de datos **comercial**	1
Base_de datos **bilingües**	1
Base_de datos **dinámica**	1
Base_de datos **explicativa**	2
Base_de datos **inteligentes**	1
Base_de datos **históricos**	1

Revisión de tarea (10)

Los adjetivos que se combinan con **base_de_datos + adjetivo** en *genómica* son:

Secuencias generadas	Total de casos
Base_de datos **disponibles**	1
Base_de datos **públicas**	2
Base_de datos **privada**	1
Base_de datos **genómicas**	2
Base_de datos **nacionales**	1
Base_de datos **norteamericanas**	1
Base_de datos **extensa**	1
Base_de datos **bibliográficas**	1
Base_de datos **genéticos**	1
Base_de datos **esencial**	1
Base_de datos **enorme**	1
Base_de datos **estructural**	1
Base_de datos **accesibles**	1

Revisión de tarea (11)

Secuencias generadas	Total de casos
Base_de datos **completa**	1
Base_de datos **relacional**	2
Base_de datos **electrónicas**	1
Base_de datos **genéticos**	1
Base_de datos **genealógica**	1
Base_de datos **poblacionales**	1

Revisión de tarea (12)

Podemos ver que la secuencia **base_de_datos** es más productiva que **base_de_conocimientos**, por lo menos en los documentos de **Bwananet** . Dado esto, ¿cuáles consideran que son sus variables dependientes e independientes? ¿Qué opinas de estos ejemplos?:

Variables dependientes:

1. Dominios de conocimiento
2. Variantes de los términos
3. Criterios de búsqueda

Variables independientes:

1. Cantidad de resultados generados
2. Criterios de constitución del corpus
3. Tipo de herramienta para búsqueda
4. Acceso a los documentos originales

Revisión de tarea (13)

Ahora, para diseñar su histograma con los datos obtenidos, ¿cómo los organizarían? Les propongo un ejemplo:

Eje vertical:

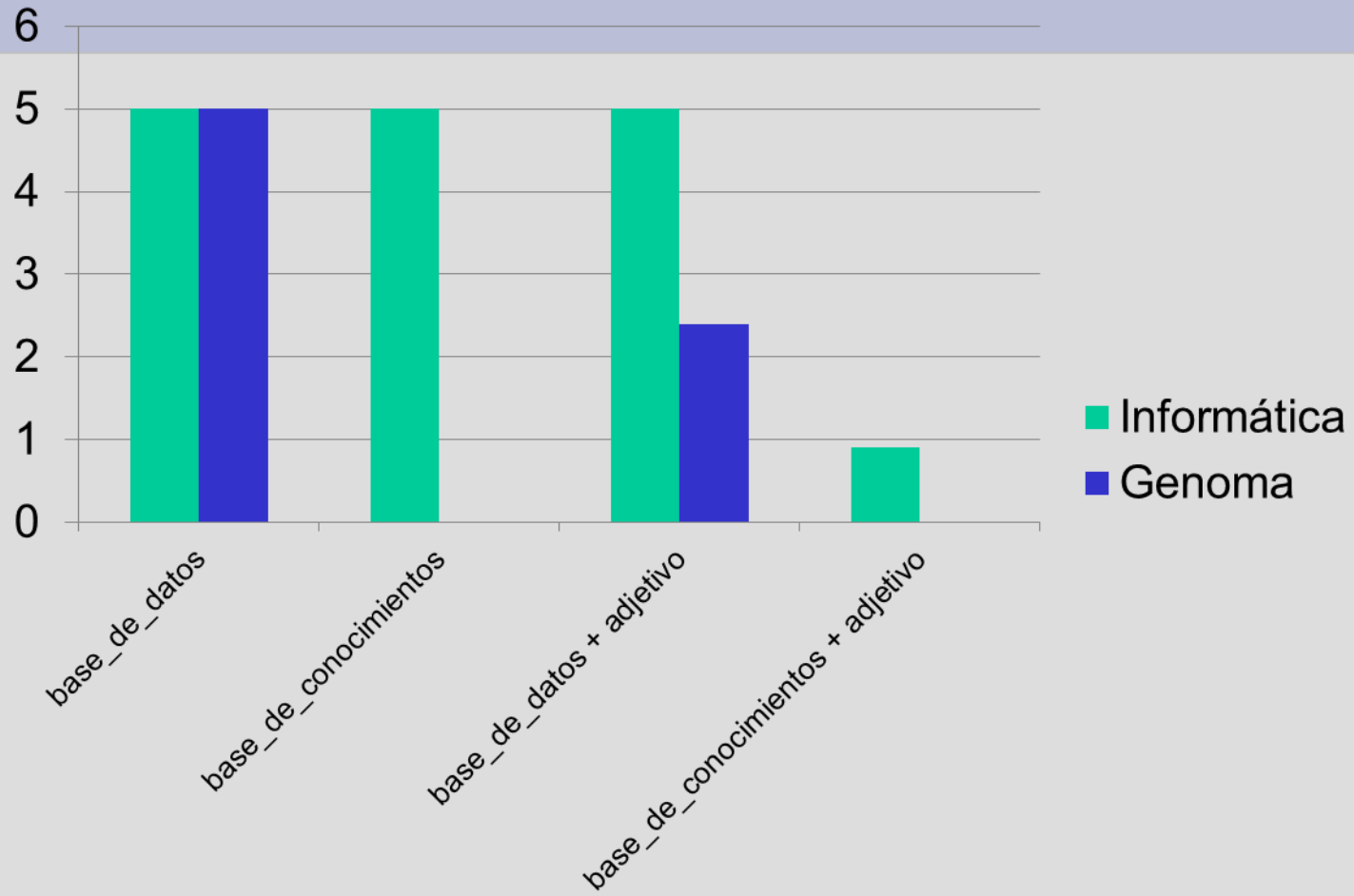
Frecuencias de uso: dado que el máximo son 50 resultados, podemos considerar que aquellas secuencias que alcanzan esta cifra son los más recurrentes. Así, les propongo una escala que siga esta secuencia: 0, 10, 20, 30, 40 y 50.

Eje horizontal:

¿Qué combinaciones fueron las que consideraron? Básicamente:

i. Las variantes **base_de_datos** *versus* **base_de_conocimiento**
ii. Las dos variantes anteriores más la inserción de adjetivo
iii. Los dominios de conocimiento: informática y genoma

Revisión de tarea (14)



Revisión de tarea (15)

Ahora, algunas preguntas para la reflexión:

1. ¿Qué otros datos podemos analizar respecto las secuencias **bases_de_datos/base_de_conocimientos?**

2. ¿Qué es lo que se puede inferir a partir del histograma anterior?

3. ¿Qué otras secuencias podrían ser derivables de las anteriores?

4. Si tuvieran que determinar las probabilidades condicionales de estas secuencias, ¿cómo lo harían?

Tablas de contingencia (1)

A partir de la búsqueda de términos en **Bwananet**, podemos darnos cuenta que requerimos algún mecanismo que nos permita organizar nuestros datos, de modo que podamos hacer cruces para distinguir distintos fenómenos. Una herramienta útil para hacer esto son las **tablas de contingencia**.

Esta clase de tablas nos ayudan a estudiar si existe alguna asociación entre una variable *fila* y otra variable *columna*. Un ejemplo sencillo es:

$X \equiv$ Se toma aspirina o placebo $(I = 2)$

$Y \equiv$ Se sufre ataque cardíaco o no $(J = 3)$.

	Ataque mortal	Ataque no mortal	No ataque
Placebo	18	171	10845
Aspirina	5	99	10933

Tablas de contingencia (2)

Una tabla de contingencia nos permite representar datos de tal forma que podemos cruzarlos y determinar si hay alguna relación (o no la hay) entre las variables que estamos analizando.

Las columnas que conforman nuestra tabla se clasifican en dos grupos: **frecuencias marginales** y **totales**. Un ejemplo es:

	Diestro	Zurdo	TOTAL
Hombre	43	9	52
Mujer	44	4	48
TOTAL	87	13	100

Las cifras en la columna de la derecha y en la fila inferior reciben el nombre de **frecuencias marginales** y la cifra situada en la esquina inferior derecha es el **gran total**.

Ejercicio (1)

A partir del siguiente ejemplo, vamos a tratar de emplear todos los cruces posibles que nos ofrece nuestra tabla de contingencias para tratar de hacer varios cálculos relacionados.

	Speeding violation in the last year	No speeding violation in the last year	Total
Car phone user	25	280	305
Not a car phone user	45	405	450
Total	70	685	755

En este caso, lo que tenemos en la tabla son cifras relacionadas con usuarios de servicio telefónico instalados en sus autos, contrastados con aquellos que no tienen este servicio, en términos de los abusos en límites de velocidad en que incurren.

Ejercicio (2)

Ahora bien, tratemos de responder a las siguientes preguntas considerando los cruces posibles.

1. ¿Cuál es la probabilidad de toparnos con una persona que sea propietaria de una línea telefónica en su auto?:

$$\mathrm{P}(\text{person is a car phone user}) =$$

Respuesta:

$$\mathrm{P}(\text{person is a car phone user}) = \frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$$

Ejercicio (3)

2. ¿Cuál es la probabilidad de toparnos con una persona que no haya cometido alguna infracción por exceder el límite de velocidad?:

$$\mathrm{P}(\text{person had no violation in the last year}) =$$

Respuesta:

$$\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$$

Ejercicio (4)

3. ¿Cuál es la probabilidad de toparnos con una persona que no haya cometido alguna infracción durante el año, y además que sea dueña de una línea telefónica en su auto? :

$$\mathrm{P}(\text{person had no violation in the last year AND was a car phone user}) =$$

Respuesta:

$$\frac{280}{755}$$

Ejercicio (5)

4. ¿Cuál es la probabilidad de toparnos con una persona que tenga teléfono en su auto, o que no haya cometido una infracción durante el último año? Esto es:

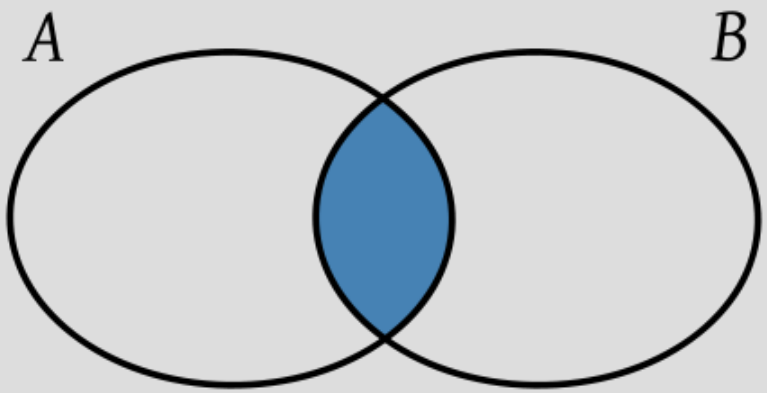
$$\mathrm{P}(\text{person is a car phone user OR person had no violation in the last year}) =$$

Respuesta:

$$\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$$

Ejercicio (6)

Si hacemos un diagrama de Venn, lo que queremos representar es:



Conjunto A: personas que tienen teléfono en su auto.

Conjunto B: personas que no han cometido una infracción.

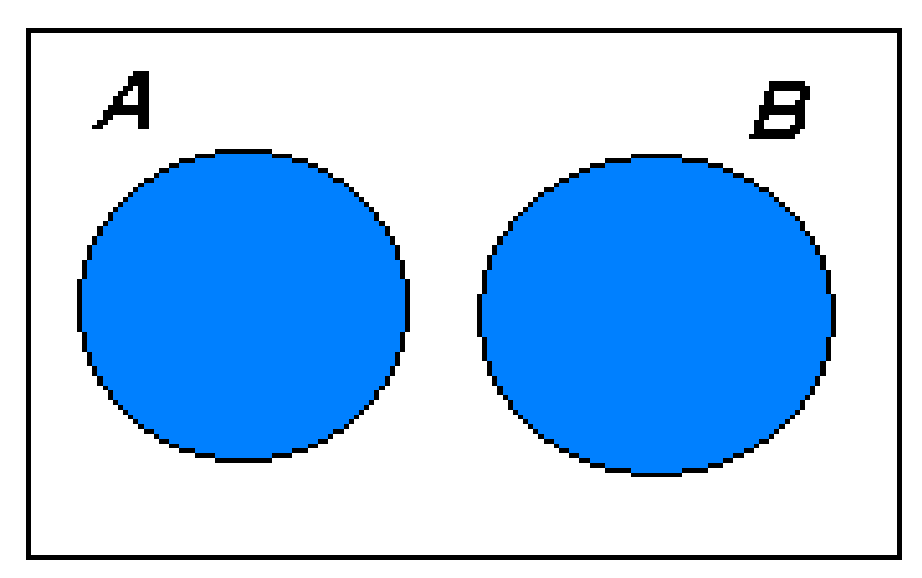
Como estos dos conjuntos no son mutuamente excluyentes, existe un traslape, ¿en dónde lo ubicamos?

Ejercicio (7)

Cuando hablamos de **eventos mutuamente excluyentes**, los podemos representar con la siguiente fórmula:

$$P(A \cap B) = P(A) + P(B)$$

Lo que equivale a:



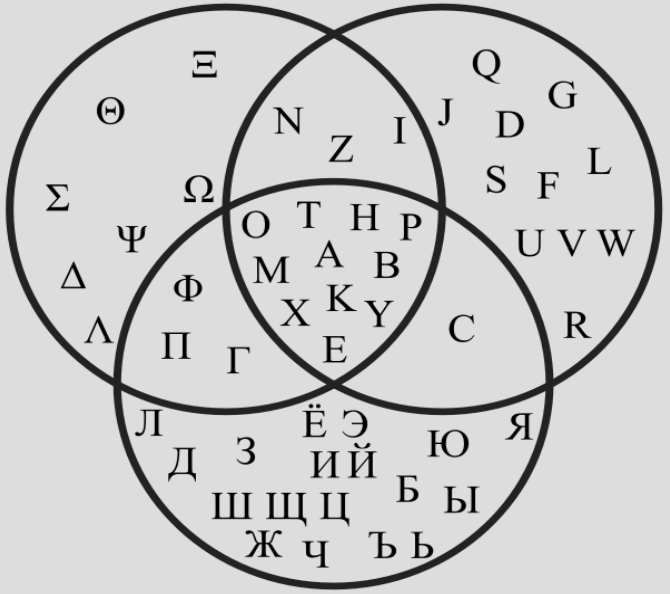
Ejercicio (8)

En contraste, la pregunta que nos plantea el ejercicio es un caso de **eventos no mutuamente excluyentes**, y se pueden expresar a través de esta fórmula:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

¿Cómo deducimos nuestras probabilidades?

Primero sumamos la probabilidad de los conjuntos A con B, y luego restamos la probabilidad de la intersección de los dos conjuntos A y B.



Ejercicio (9)

5. ¿Cuál es la probabilidad de que una persona tenga un teléfono en su auto, dada la posibilidad de que tenga una infracción durante el año? :

$$\mathrm{P}(\text{person is a car phone user GIVEN person had a violation in the last year}) =$$

Respuesta:

$$\frac{25}{70}$$

Nuestro espacio de búsqueda se reduce solamente a aquellas personas que hayan cometido una infracción.

Ejercicio (10)

5. ¿Cuál es la probabilidad de que una persona no tenga un teléfono en su auto, dada la posibilidad de que no tenga una infracción durante el año? :

$$\mathrm{P}(\text{person had no violation last year GIVEN person was not a car phone user}) =$$

Respuesta:

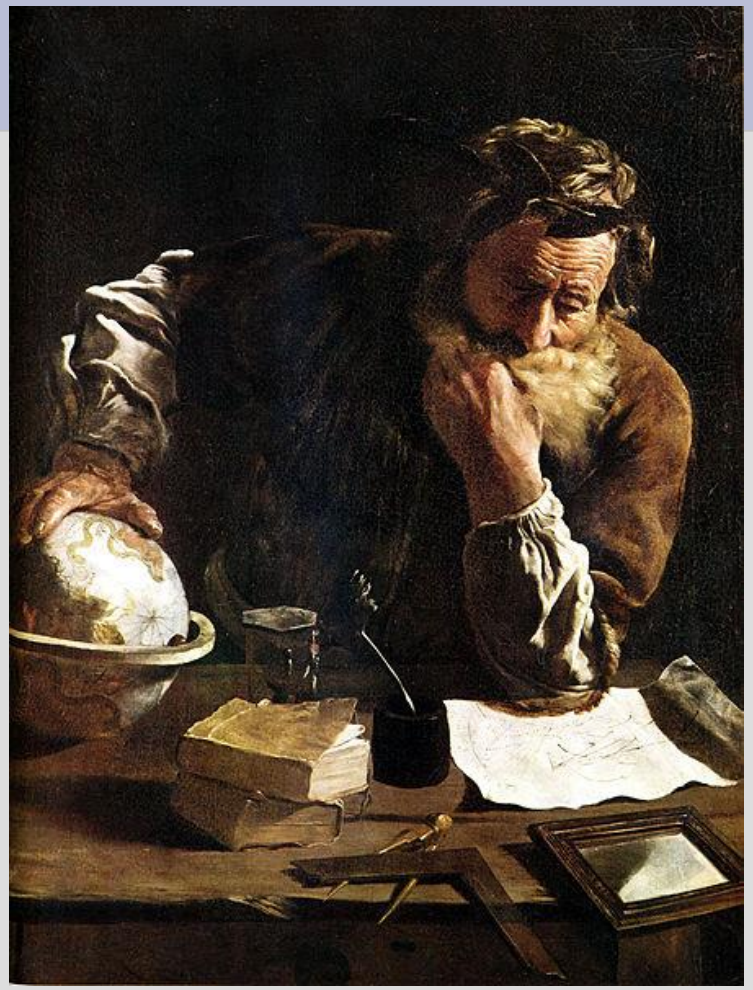
$$\frac{405}{450}$$

Nuestro espacio de búsqueda se reduce solamente a aquellas personas que no cuentan con un teléfono en sus autos.

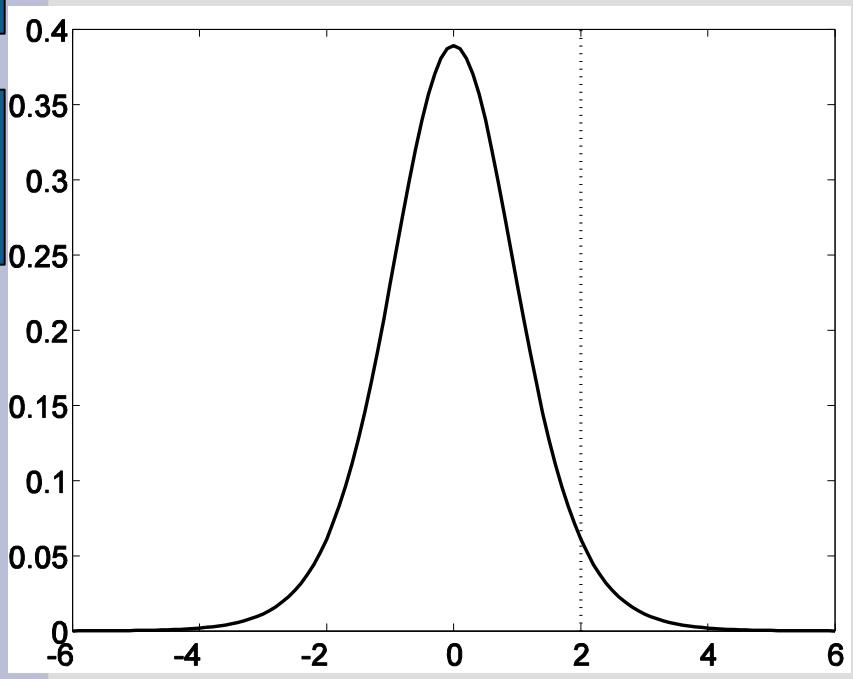
Buscando causas y efectos (1)

Las tablas de contingencias, de acuerdo con el ejercicio que hicimos, nos sirven como una especie de *calculadora* que nos permite delimitar nuestras búsquedas a un conjunto de datos específicos, los cuales nos ofrecen una representación numérica del fenómeno (o los fenómenos que estamos estudiando).

La idea de esto es simple: lo que tratamos es entender y describir el comportamiento de una enorme población a partir de una muestra (la cual pretendemos que esté lo mejor seleccionada y regulada posible), aplicando todos los cruces que podamos a dicha muestra. A esta tarea la denominamos **inferencia estadística.**



Buscando causas y efectos (2)



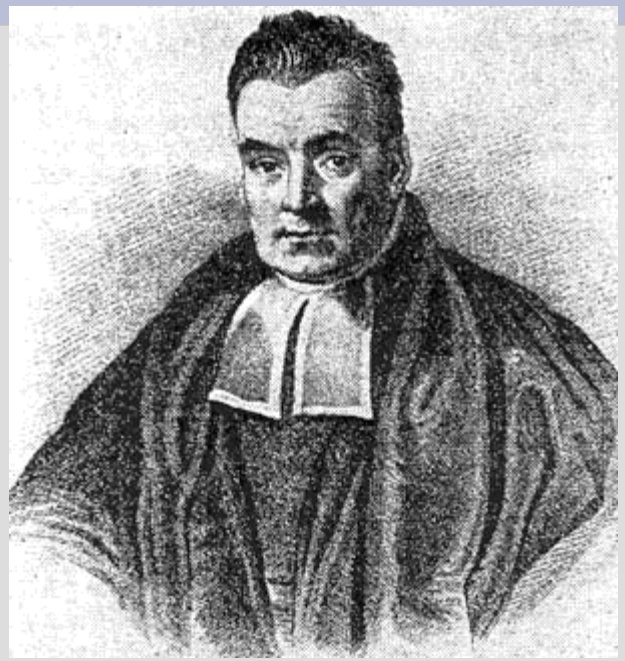
Entrando en detalles, una inferencia estadística es una técnica que nos ayuda a delimitar y clarificar los resultados que obtenemos de nuestro análisis, estableciendo siempre un margen de error (lo que nos da pie a realizar ajustes, si esto es necesario).

Estos márgenes de error no necesariamente deben ser visto como “fallas”, sino más bien como una gradación que nos ayuda a determinar cuál es el peso real que tienen los factores que consideramos involucrados en el fenómeno estudiado.

Buscando causas y efectos (3)

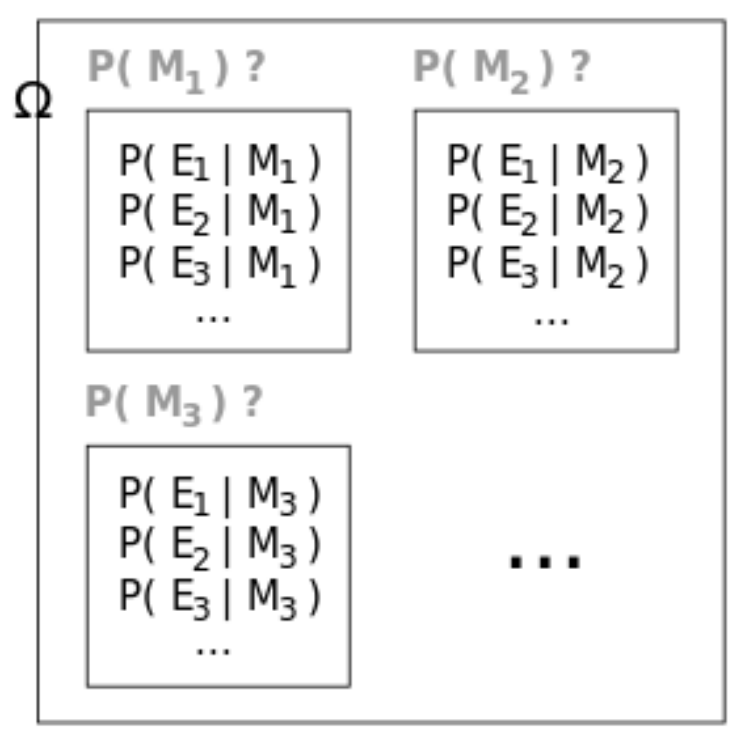
Un buen ejemplo de cómo operan las inferencias estadísticas es el **teorema bayesiano**, planteado por **Thomas Bayes** (1701-1761).

A grandes rasgos, una inferencia bayesiana considera que las evidencias u observaciones son una vía fundamental para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta (o refutable, si es el caso).



Thomas Bayes

Buscando causas y efectos (4)



Lo anterior podemos entenderlo del siguiente modo: supongamos que tenemos una hipótesis que nos sirve para explicar un fenómeno dado, y la contrastamos con su negación. Para resolver este conflicto, vamos recolectando evidencia que se considera consistente o inconsistente con una alguna de nuestras dos hipótesis.

A medida que la evidencia se acumula, el grado de creencia en una hipótesis se va modificando. De este modo, las hipótesis con un grado de creencia muy alto se pueden tomar como verdaderas, mientras que las que tienen un bajo grado de creencia muy bajo son vistas como falsas.

Buscando causas y efectos (5)

Veamos un ejemplo para entender mejor cómo operan una inferencia bayesiana: ¿qué factores influyen para que una persona enferme de cáncer en los pulmones?

(a) Fumar

(b) Respirar aire contaminado

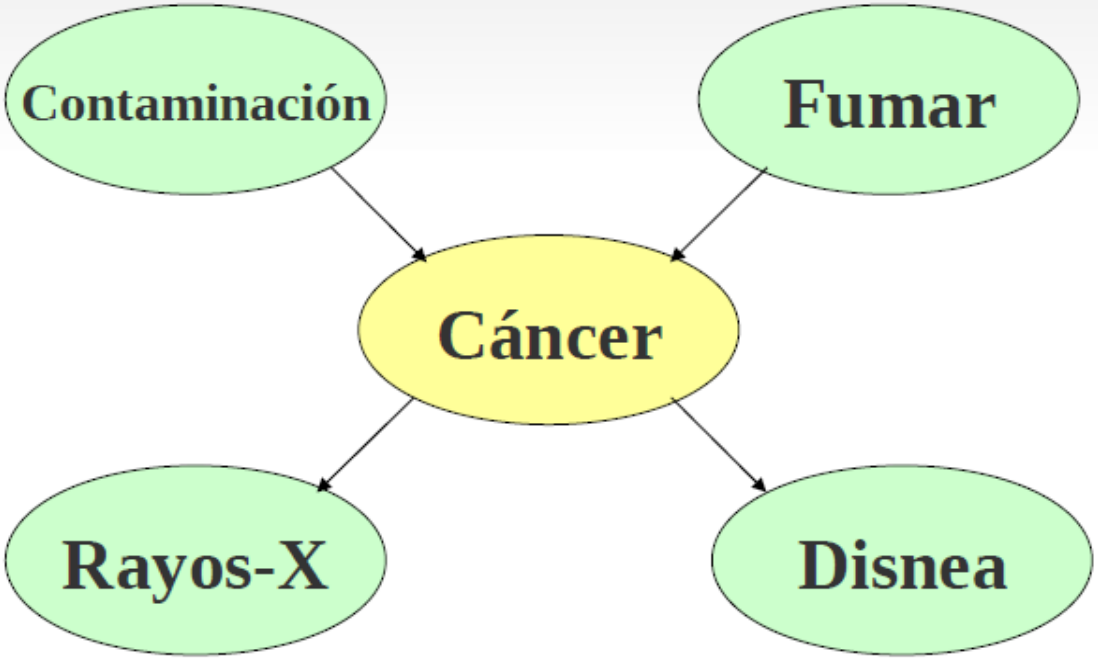
Una vez resuelta esta duda, hagamos otra pregunta: ¿cuál es la mejor forma de detectarlo?

(i) A través de una radiografía

(ii) Detectando síntomas de disnea (esto es, dificultades para respirar)

Buscando causas y efectos (6)

Tenemos entonces dos posibles causas (contaminación *vs.* fumar), un efecto (contraer cáncer), y dos maneras de detectarlo (vía rayos-X *vs.* Padecer de disnea).



Preguntas maliciosas:

1. ¿Cuál sería el peor paciente posible?

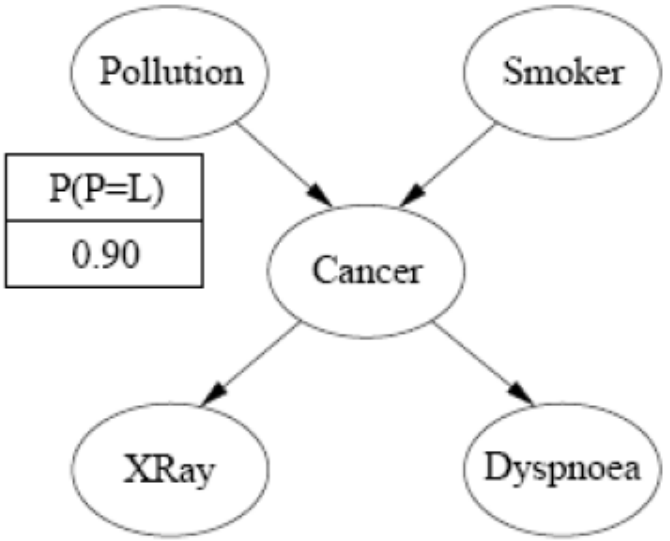
2. Si fumo, pero no sufro disnea, y los rayos X no me detectan nada en el pulmón, ¿puedo llegar a sufrir cáncer?

3. ¿Todos los habitantes del DF, por el hecho de estar expuestos a la contaminación, pueden sufrir cáncer?

Buscando causas y efectos (7)

Si se dan cuenta, el problema no es decidir qué causas son más nocivas para contraer cáncer, o cuál sea la manera más efectiva para detectarlo. El punto es más bien:

Considerando todos estos factores como posibles, ¿cuál es la combinación que da el peor escenario posible?
Ahora, si vamos reduciendo nuestras secuencias de factores, ¿podemos llegar a un diagnóstico preciso?
Incluso, si comparamos con un gran número de cuadros clínicos, ¿qué factores son los más recurrentes?



P(S=T)
0.30

P(P=L)
0.90

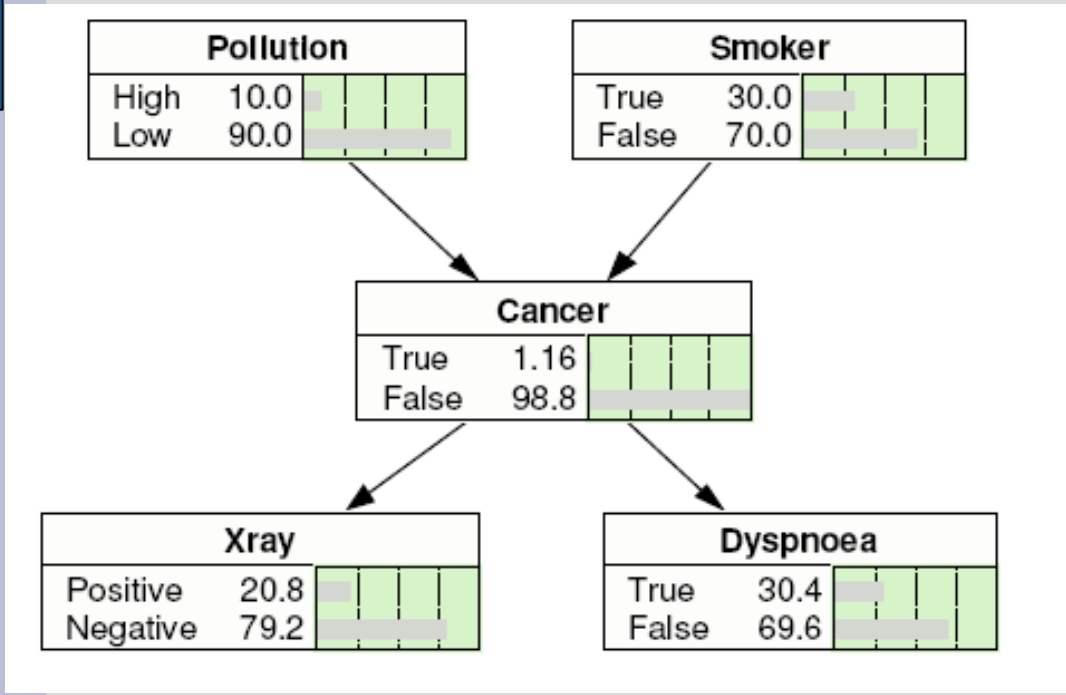
P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

Buscando causas y efectos (8)

Siguiendo este razonamiento, parece que en un 30% de los casos de fumadores, ésta puede contraer cáncer, y podemos detectarlo en un 30.4% de los casos considerando si tienen problemas para respirar.



Imaginemos que la combinación de estos dos factores nos da lo siguiente:

1. Fumar CAUSA cáncer.

2. La disnea ES-UN EFECTO del cáncer.

Si evito fumar, puede ayudar a que no padezca disnea, y la combinación de ambos hace que tenga pocas probabilidades de padecer cáncer.

Buscando causas y efectos (9)

Nuevamente, partimos de lo que hemos visto sobre probabilidades condicionales: lo que queremos establecer es si en una cadena de eventos dada, ¿podemos reconocer cuáles son dependientes entre sí, y cuáles son marginales? Esto podemos determinarlo con la siguiente fórmula:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)}$$

Esto quiere decir: “sea {A1, A2, A3...An} un conjunto de sucesos mutuamente excluyentes y exhaustivos, tales que la probabilidad de cada uno de ellos es distinta de cero”.

A esto le añadirmos: “sea B un suceso cualquiera del que se conocen las probabilidades condicionales P(B | Ai)”.

Al final, siguiendo la fórmula, deducimos:

- $P(A_i)$ son las probabilidades a priori.
- $P(B \mid A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i \mid B)$ son las probabilidades a posteriori.

Esto se cumple $\forall i = 1 \ldots n$

Buscando causas y efectos (10)

Para mayores detalles, este tipo de inferencias se basan en el **teorema de Bayes**, el cual nos dice que:

Sea $\{A_1, A_2, A_3 \ldots A_i \ldots A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B \mid A_1)$. Entonces, la probabilidad $P(A_1 \mid B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i|B)$ son las probabilidades a posteriori.

Buscando causas y efectos (11)

Si les interesa profundizar en la forma en como se propone y desarrolla el teorema de Bayes, pueden revisar la entrada que la *Stanford Encyclopaedia of Philosophy* le dedica al tema. La liga es:

http://plato.stanford.edu/entries/bayes-theorem/

Pregunta: como mera reflexión, ¿qué aplicaciones le darían al teorema de Bayes, si con él quisieran analizar un fenómeno lingüístico? ¿Qué podrían estudiar?

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/meacutetodos-y-teacutecnicas-de-investigacioacuten-cuantitativa.html