PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos y técnicas de investigación cuantitativa

César Antonio Aguilar

Facultad de Lenguas y Letras

25/03/2013

Cesar.Aguilar72@gmail.com

Algunos conceptos básicos (1)

En esta sesión, y basándonos en el capítulo 1 del libro de Christopher Butler, *Statistics in Linguistics*, vamos a revisar algunos conceptos básicos, en concreto:

1. Poblaciones
2. Azar
3. Muestras aleatorias
4. Tipos de variables: dependientes e independientes

Nota:

Pero antes de ver estos conceptos, revisemos lo que quedó como tarea la clase pasada…

Tarea (1)

La clase pasada tratamos de establecer una serie de probabilidades para determinar qué tipo de clítico **se** teníamos en 4 casos: forma impersonal, forma pasiva, voz media y forma reflexiva. Recordemos los ejemplos:

1. *Se* como marca de forma impersonal: ***se*** *venden casas.*

2. *Se* como marca de media pasiva: *una computadora* ***se*** *define como una máquina electrónica para hacer cálculos.*

3. *Se* como marca de voz media: ***se*** *me hizo tarde.*

4. *Se* como marca reflexividad: ***se*** *miró al espejo.*

Tarea (2)

Si recuerdan, establecimos algunas condiciones para delinear el comportamiento que asume este clítico en estas 4 formas. Así:

1. Si este clítico se encuentra al inicio de una oración, da lugar a una forma impersonal.

2. Si no es el caso, entonces tenemos una forma en pasiva.

3. Si es una forma pasiva, entonces no debe ir asociada a un verbo reflexivo.

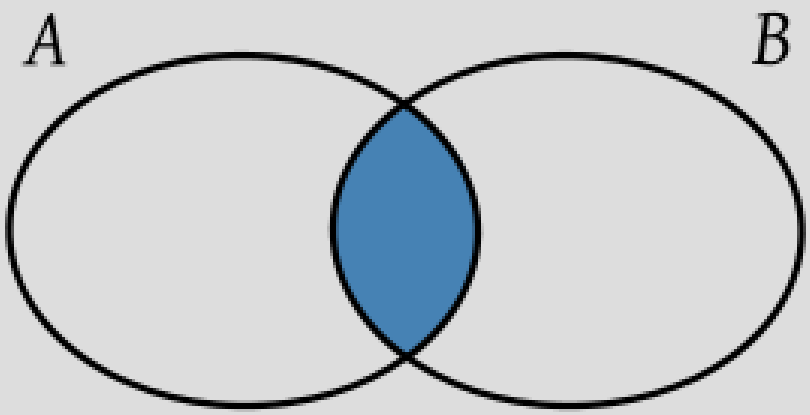
4. Si no es el caso, entonces tenemos un clítico **se** en una forma reflexiva.

Tarea (3)

En la clase pasada, establecimos 3 teoremas o fórmulas que nos ayudaban a analizar y predecir el comportamiento de **se**, basándonos en los axiomas anteriores.

Así, consideramos que una relación propia de una teoría de conjuntos, en este caso la **intersección**, en donde suponemos que estas formas gramaticales se dan cuando **se** establece una relación con un elemento concreto de la oración.

La idea aquí es: supongamos que el conjunto A agrupa las posiciones posibles del clítico **se** en una oración, y B representa los elementos posibles que, al asociarse con A, generan una de las formas que hemos considerado. Lo que buscamos entonces es dar con todos los elementos que se agrupan en torno al sub-conjunto $A \cap B$.



Tarea (4)

Tenemos entonces los teoremas siguientes. Para las construcciones en forma impersonal:

$$P(Pos_{Suj} \mid Se_{imp}) = P(Pos_{Suj} \cap Se_{imp}) / P(Se_{imp})$$

Para las construcciones en forma pasiva:

$$P(Se_{pas} \mid V_{no_reflex}) = P(Se_{pas} \cap V_{no_reflex}) / P(Se_{pas})$$

Para las construcciones en forma reflexiva:

$$P(Se_{reflex} \mid V_{reflex}) = P(Se_{reflex} \cap V_{reflex}) / P(Se_{reflex})$$

Tarea (5)

Volvamos a nuestra propuesta final: ¿Cuáles son las probabilidades de toparnos con una construcción que tenga un clítico **se inserto**? Veamos:

1. Total de casos con posibles con clítico SE: 100%
2. Total de casos en donde **se** puede proyectar formas impersonales: 50%
3. Total de posibles casos en pasiva: 25%
4. Total de posibles casos en forma reflexiva, dado el tipo de verbo de la oración: 25%

Aclaración: ¿qué estamos tomando en cuenta en nuestro cálculo?:

1. La posición inicial/no inicial de **SE** en una oración.
2. La relación que puede o no tener con un potencial verbo reflexivo.
3. Según nuestro pronóstico, si el clítico aparece en posición inicial, es **impersonal**, si no es el caso, y aparece relacionado con un verbo reflexivo, es **reflexivo**. Si no es ninguno de los dos previos, es **pasivo.**

Tarea (6)

Aplicando los criterios anteriores, si quisiéramos deducir cuál es el comportamiento de la voz media y hacer su predicción, ¿qué tenemos que considerar?

1. ¿La posición del clítico en la oración?
2. ¿La presencia/ausencia de un pronombre acusativo/dativo unicado en una posición inmediata a **SE**?
3. ¿Ambos factores?

Respuesta: como estamos trabajando con probabilidades, vamos a tomar la opción 3, esto es, combinamos ambos factores.

Tarea (7)

Vamos entonces a considerar estas dos opciones. Supongamos entonces que las construcciones en voz media proyectan al clítico **SE** en posición inicial, igual que la forma impersonal.

Empero, las formas impersonales deben diferenciarse de las formas medias en el siguiente rasgo: éstas requieren la presencia de un pronombre acusativo/dativo para asumir el formato de voz media. Veamos:

[Se le [hizo tarde [para ir al trabajo PREP”] V”] FLEX”]

Tarea (8)

Si ya tenemos el problema para resolver la posición inicial, simplemente lo volvemos a ocupar:

$$P\,(Pos_{Suj} \mid Se_{imp}) = P(Pos_{Suj} \cap Se_{imp}) \,/\, P(Se_{imp})$$

Ahora, formulemos nuestra sospecha de la siguiente manera:

$$P\,(Se_{vm} \mid PRO_{acus/dat}) = P(Se_{vm} \cap PRO_{acus/dat}) \,/\, P(Se_{vm})$$

En otras palabras, esto quiere decir que: *la probabilidad de que un clítico SE aparezca asociado a un pronombre acusativo o dativo, es igual a la probabilidad de que ambos aparezcan ligados, entre la probabilidad de que tal clítico marque una voz media.*

Tarea (9)

¿Cuál es nuestro porcentaje de probabilidades? Parece que es fácil de deducir:

1. Total de casos con posibles con clítico SE: 100%
2. Total de casos en donde **se** pueden ir en posición inicial:50%
3. Total de casos que no van en posición inicial: 50%
4. Total de casos que pueden proyectar formas impersonales: 25%
5. Total de posibles casos que no son impersonales, porque van asociados con un pronombre acusativo/dativo: 25%
6. Total de posibles casos en pasiva: 25%
7. Total de posibles casos en forma reflexiva, dado el tipo de verbo de la oración: 25%

Observación (1)

Es posible que nuestro pronóstico sea conservador, porque no estamos considerando excepciones en donde tengamos combinaciones tales como:

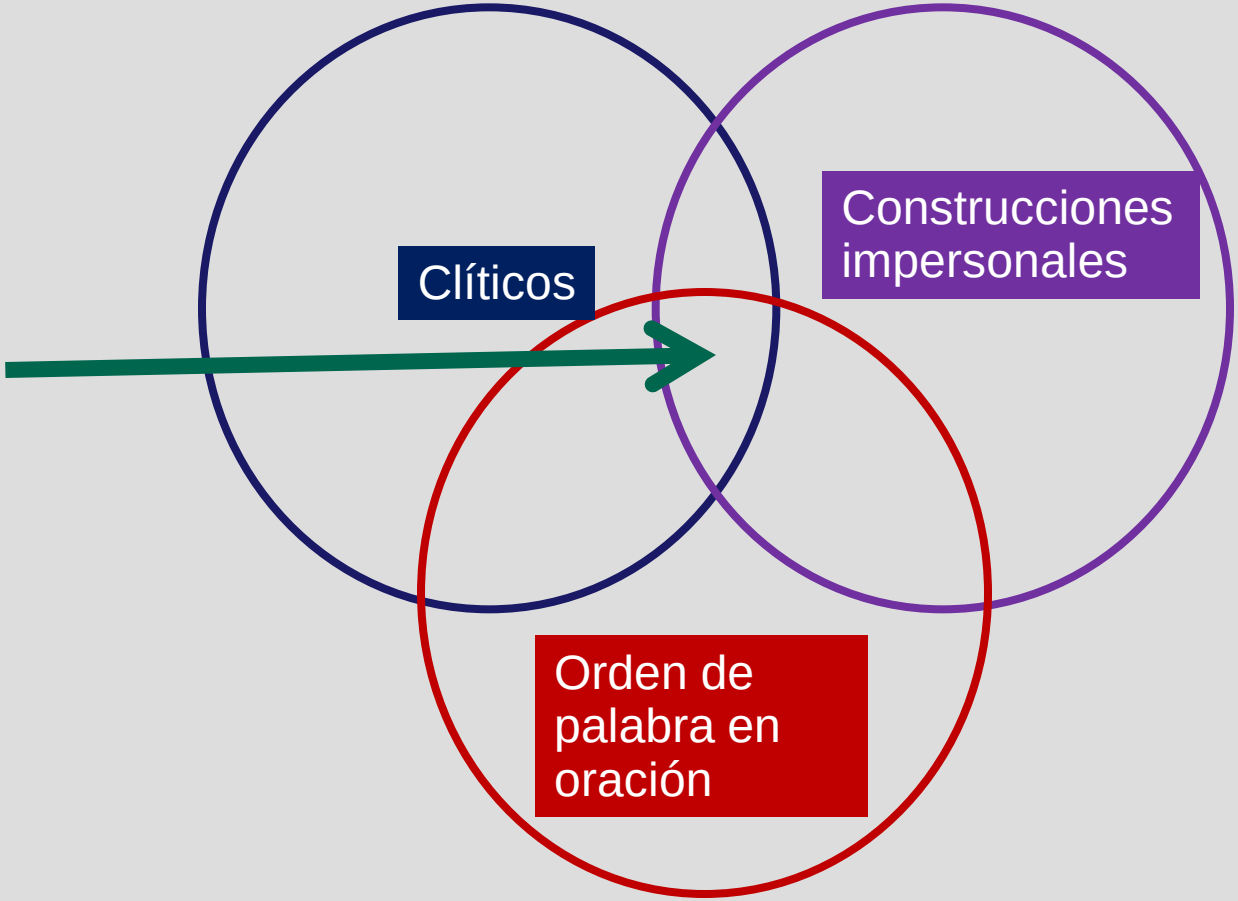
1. Se les vendió una casa en un precio razonable.
2. Se lo pensaron mucho antes de comprarla.
3. Indíque**sele** hacia dónde va

Habría que contar estos posibles casos y hacer reajustes. Con todo, parecen casos que no afectan mucho nuestros pronósticos, considerando que tenemos cubiertos al menos 4 casos sumamente recurrentes en español.

Observación (2)

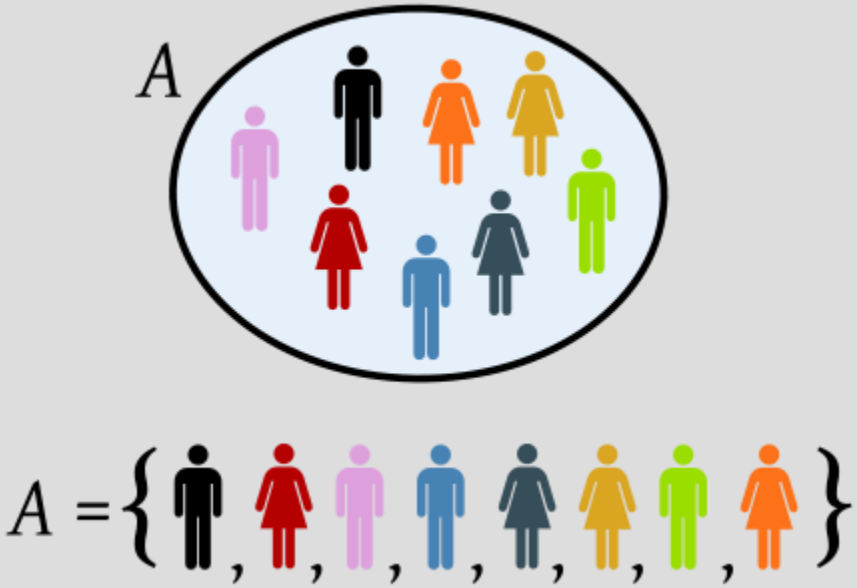
Ahora, volviendo a la lógica de lo que estamos haciendo, es determinar cuán probable es que los elementos de varios conjuntos mantienen relaciones de intersección entre sí, de forma tal que esto nos ayuda a confirmar que un fenómeno es imposible de ocurrir, si no tomamos en cuenta tales relaciones.

Construcciones impersonales con el clítico SE en español



¿Qué es una población? (1)

Ahora bien, el cálculo que estamos haciendo es una primera aproximación, basada en una serie de supuestos que observamos a través de un conjunto de casos seleccionados al azar. Podríamos decir entonces que estos casos son parte de un **conjunto**, esto es, de una colección de casos que comparten rasgos similares. Un ejemplo sencillo, pensemos en un conjunto de personas:



Lo que tenemos aquí es un conjunto de 8 personas, las cuales comparten una serie de rasgos comunes, las cuales se integran a este conjunto sin pensar demasiado en particularidades tales como su edad, su sexo, su condición económica, etc.

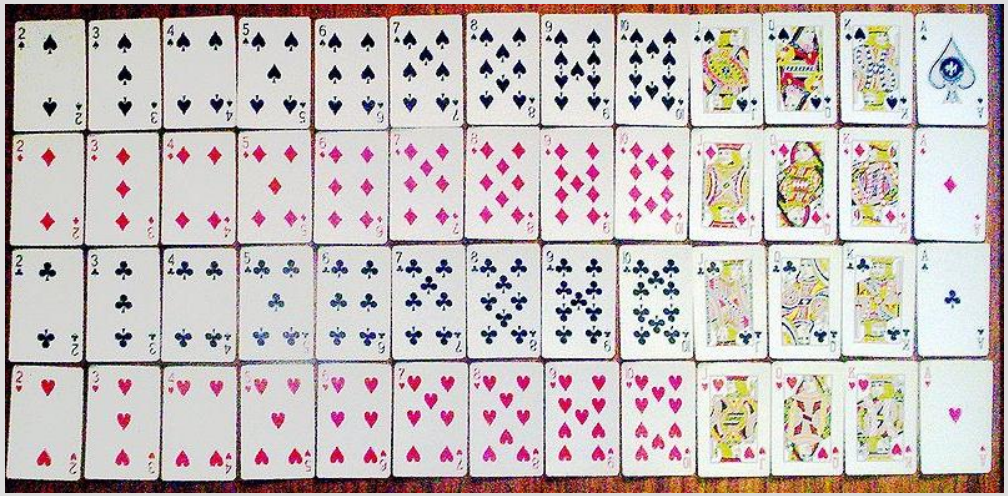
¿Qué es una población? (2)

Sin embargo, este conjunto no está del todo seleccionado de forma arbitraria: si suponemos que vamos a formar un conjunto, debe de haber un mínimo de rasgos que compartan los elementos que lo conformen. A esta relación se le conoce como **pertenencia**, y se representa con el símbolo $\in$. Otro ejemplo: supongamos que formamos un conjunto de polígonos, y excluimos entonces todas las figuras bidimensionales que no lo sean:

Analizando poblaciones (1)

Hasta aquí, podemos considerar que una población es equivalente aun conjunto, esto es, a una colección de miembros u objetos que comparten ciertos rasgos.

Sin embargo, si nos preguntáramos: ¿cuál es el comportamiento que muestran los miembros de ese conjunto?, de entrada parece ser que no nos basta con enumerar todos sus miembros, p.e.:



- Picas ♠,
- Corazones ♥,
- Diamantes ♦, y
- Tréboles ♣.

La baraja inglesa maneja dos colores o palos (in. *suit*): negro y rojo. Maneja 4 figuras:

Analizando poblaciones (2)

Claro, si aplicamos algunas relaciones propias de la teoría de conjuntos, podemos llegar a hacer una descripción muy general, p.e., el número total de combinaciones que pueden establecer los subconjuntos de palos y rangos, aplicando un **producto cartesiano**:

El producto cartesiano del conjunto de trece rangos de la baraja inglesa

$$Rangos = \{As, K, Q, J, 10, 9, 8, 7, 6, 5, 4, 3, 2\}$$

con el de los cuatro palos:

$$Palos = \{\spadesuit, \heartsuit, \diamondsuit, \clubsuit\}$$

conjunto de las 52 cartas de la baraja:

$$Baraja = \{(As, \spadesuit), (As, \heartsuit), (As, \diamondsuit), \ \ldots \ , (2, \heartsuit), (2, \diamondsuit), (2, \clubsuit)\}$$

Esto se puede sintetizar del siguiente modo:

$$Baraja = Rangos \times Palos = \{(x, y) \mid x \in Rangos \ \wedge \ y \in Palos\}$$

Analizando poblaciones (3)

Sólo para no olvidarlo, la fórmula general para desarrollar un producto cartesiano es:

El **producto cartesiano** de *A* y *B* es el conjunto *A* × *B* cuyos elementos son los pares ordenados (*a*, *b*), donde *a* es un elemento de *A* y *b* un elemento de *B*:

$$A \times B = \{(a, b) : a \in A \text{ y } b \in B\}$$

En resumen, cada carta representa esta relación generada por el producto cartesiano: a toda secuencia de números o rangos le corresponde un color y un palo (esto es, una figura).

¿Y cómo se juega con esto? Por ejemplo si jugamos póker, lo que buscamos es establecer una secuencia de rangos y palos alta, en un N número de jugadas.

Analizando poblaciones (4)

Ya metidos en el juego del póker, para ganar justo lo que nos interesa son las combinaciones de rangos y palos más altas, de modo que en un número limitado de jugadas (o *manos*), debemos fijar una secuencia con estos valores, la cual sea ganadora. Si la matemática no nos falla, tenemos 10 tipos de jugadas, con sus posibles combinaciones:

	Jugada	*Jugada en inglés*	Descripción	Ejemplo	Combinaciones Posibles
1	**Escalera real** o **flor imperial**	*Royal flush*	Cinco cartas seguidas del mismo palo del 10 al As.	A♣ K♣ Q♣ J♣ 10♣	4 entre 2.598.960
2	**Escalera de color**	*Straight flush*	Cinco cartas consecutivas del mismo palo.	7♦ 8♦ 9♦ 10♦ J♦	36 entre 2.598.960
3	**Póker**	*Four of a kind* o *Quad*	Cuatro cartas iguales en su valor.	9♥ 9♠ 9♣ 9♦ 3♦	624 entre 2.598.960
4	**Full**	*Full House*	Tres cartas iguales en su valor(trío), más otras dos diferentes en su valor(par).	6♦ 6♠ 6♥ 3♣ 3♠	3.744 entre 2.598.960
5	**Color**	*Flush*	Cinco cartas del mismo color (palo), sin ser necesariamente consecutivas.	2♥ 7♥ J♥ A♥ 4♥	5.108 entre 2.598.960
6	**Escalera**	*Straight*	Cinco cartas consecutivas sin importar el palo.	3♦ 4♣ 5♦ 6♠ 7♥	10.200 entre 2.598.960
7	**Trío**	*Three of a kind* o *Set*	Tres cartas iguales de valor.	8♥ 8♠ 8♣ 2♣ 10♦	54.912 entre 2.598.960
8	**Dobles parejas**	*Two pair* o *Pocket*	Dos pares de cartas del mismo número (par y par).	Q♦ Q♠ 5♥ 2♣ 5♠	123.552 entre 2.598.960
9	**Pareja**	*One pair*	Dos cartas iguales de número (y tres diferentes).	K♣ K♦ 7♣ 2♠ J♥	1.098.240 entre 2.598.960
10	**Carta alta**	*High card*	Gana quien tiene la carta más alta de todas.	Q♦ 7♠ 5♥ 3♣ 10♠	1.302.540 entre 2.598.960

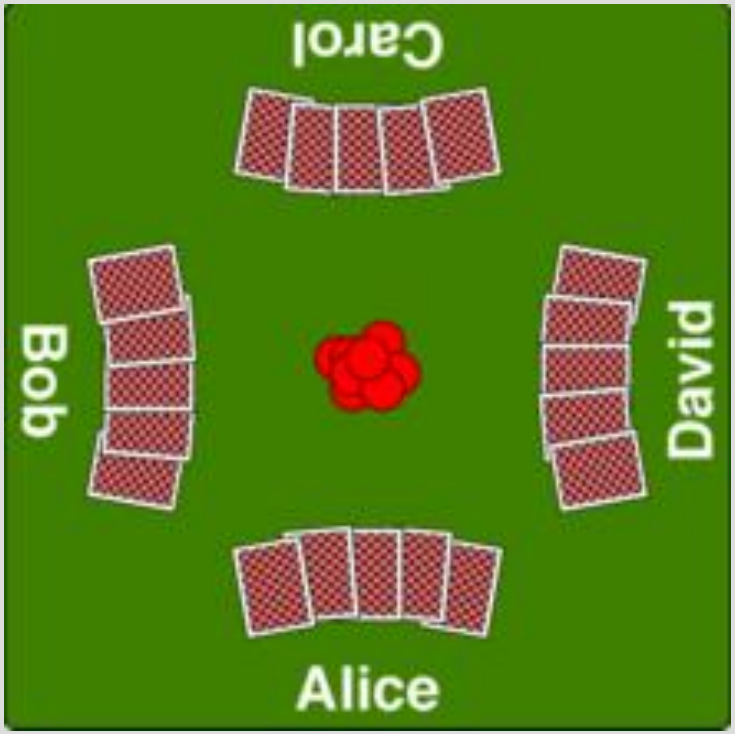
Analizando poblaciones (5)

Como jugadores, no nos interesa tanto saber todas estas cosas sobre la baraja y el juego de póker, sino cosas más *prácticas* como:

1. Si tengo un par de Reinas en una mano, ¿cuántas necesito para completar un póker de 4 Reinas?

2. ¿Lo puedo hacer en una jugada, máximo?

3. Si ya salió una Reina, pero quedan dos comodines en la baraja todavía, ¿es posible que me toque uno en la siguiente jugada?



¿Qué es el azar? (1)

Resolver a preguntas como las anteriores, requiere que introduzcamos una causa que rige en buena parte de lo que podemos considerar **fenómenos complejos** (incluido el póker y el lenguaje natural): el **azar**.

El azar lo podemos entender como una condición que, según nuestro razonamiento, puede o no afectar una correlación de dos (o más) variables estadísticas correlacionadas.

En caso de que las afecte, entonces podemos considerar que no tenemos buenos argumentos para explicar por qué se da esta relación.

En cambio, si el azar —a pesar de su manifestación— no afecta esta correlación entre las variables, entonces quiere decir que tal correlación está determinada (y de manera directa) por la presencia (o ausencia) de dichas variables.

¿Qué es el azar? (2)

Hay muchísimos ejemplos en la vida cotidiana que tienen que ver con el azar, el más típico: arrojar una moneda al aire, y adivinar que el lado de la moneda que se va a ver cuando caiga al piso es cara (en español mexicano: *Sol*), o cruz (en español mexicano: *Águila*).

Hablando *derechamente*: lo que vamos a hacer es jugar a los *volados* con un *Peso*, que no esté *calado*, y que sus dos lados sean un *Águila* y un *Sol* nomás a dos opciones: águila y/o sol.

¿Qué es el azar? (3)

Traduciendo esto a un lenguaje más técnico, lo que queremos hacer es:

1. Determinar si una secuencia de eventos, fijada de antemano (una moneda que nos muestre uno de sus lados, *Águila* o *Sol*), estén afectados por el azar o no.

2. Un *espacio de muestreo* que nos permita acomodar tal secuencia de eventos para hacer nuestras pruebas.

3. Una *medida de probabilidad* para calcular cuán dependiente o independiente es la relación que hay entre los factores que configuran los eventos a analizar.

Traducción: nuestra secuencia de eventos es echarnos varios *volados* con una moneda. Ahora, nuestro espacio de búsqueda se reduce a dos opciones: águila y/o sol. Finalmente, nuestra medida de probabilidad es la cantidad de *volados* para que nos salga una de las dos opciones que tenemos.

¿Qué es el azar? (4)

Para deducir nuestras medidas de probabilidad, decimos que:

Tenemos dos opciones, a las cuales les damos un valor de 0.5:

Probabilidad de que caiga Águila = P({Ag}) = 0.5
Probabilidad de que caiga Sol = P({So}) = 0.5
La suma de estas dos probabilidades nos da un total de 1, que es el valor de máximo de nuestra secuencia de eventos = P({Ag,So}) = P({Ag})+P({So}) = 1

Podemos describir el comportamiento de nuestro evento a partir de axiomas:

$0 \leq P(A) \leq 1$ para cada evento $A \subseteq \Omega$
$P(\Omega) = 1$
$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Esto equivale a decir que el valor de la probabilidad de un evento A, es decir P(A) debe estar en un rango que va de 0 a 1. Así, la probabilidad de que ocurra al menos un evento del espacio muestral Ω es de 1. Finalmente, la probabilidad de que ocurra uno u otro es la suma de las probabilidades, es decir, la probabilidad de que ocurra águila o sol en el lanzamiento de una moneda es 1.

¿Qué es el azar? (5)

Ahora, lancemos 3 veces nuestra moneda al aire. Las combinaciones que se den en esos tres lanzamientos conforma nuestro espacio de muestreo (*Es*). Podemos representar estos lanzamientos del siguiente modo:

- *Es* ={AgAgAg; AgAgSo; AgSoAg; SoAgAg; AgSoSo; SoAgSo; SoSoAg; SoSoSo}

- Asumimos que cada hecho es *equiprobable*.
- **Pregunta:** ¿cuántas probabilidades tenemos de que caiga 2 veces Sol?
- A = {SoSoAg; SoAgSo; AgSoSo}
- P(A) = P({SoSoAg}) + P({SoAgSo}) + P({SoAgSo})
- = 1/8 + 1/8 + 1/8
- =3/8

¿Qué es el azar? (6)

Una pregunta no tan ingenua: ¿Y si nuestra moneda esta trucada y no permite que caiga Sol mas que una sola vez? Nuestras opciones de combinación se reducen, por ejemplo:

Es ={AgAgAg; **AgAgSo; AgSoAg; SoAgAg**; AgSoSo; SoAgSo; SoSoAg; SoSoSo}

Pregunta: ¿Cuantas probabilidades tenemos de que caiga Sol por lo menos una sola vez?

B = {AgAgSo; AgSoAg; SoAgAg}

P(B) = P({AgAgSo}) + P({AgSoAg}) + P({SoAgAg})

= 2/8 + 2/8 + 2/8

=6/8

¿Qué es el azar? (6)

Podemos graficar nuestros intentos en la siguiente tabla:

Águila

Ag								
Ag								
Ag								
Ag			2/8					
Ag		2/8						
Ag	2/8							
Ag								
Ag	So	So	So	So	So	So	So	So

Sol

Probabilidades independientes y condicionales (1)

Para nuestros análisis, vamos a tomar en cuenta dos tipos de probabilidades: una en donde suponemos que no hay una relación de ningún tipo entre dos (o más variables) para que se dé un fenómeno. Decimos que tales fenómenos están regidos por el azar (y de manera absoluta).

Algo tan simple como decir: *me duelen las muelas del juicio, seguro es que va a llover hoy*, es postular justo un fenómeno que, si se llegara a cumplir, no veríamos una relación clara entre las variables que intervienen en él. La fórmula para representar esto es:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Probabilidades independientes y condicionales (2)

En cambio, si hay una relación subyacente entre dos variables, y ésta condiciona la manifestación de un fenómeno, entonces hablamos de **probabilidades condicionales**. En otras palabras, lo que queremos señalar es que cada vez que se presenta una asociación entre dos variables (A y B), observamos una alta probabilidad de que ocurra un fenómeno. Así, las dos variables se influyen mutuamente. La fórmula para presentar esto es:

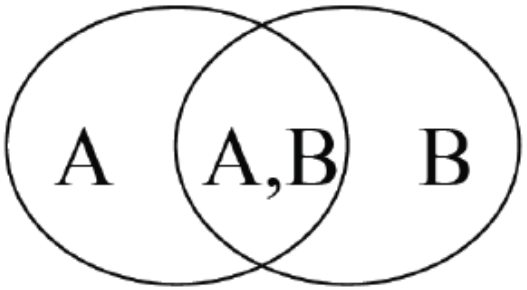
$$p(B|A) = p(A \cap B) / p(A)$$

Volviendo a nuestro juego de cartas, podríamos decir que conforme resolvamos quién tiene las Reinas que nos faltan para nuestro póker, eso condiciona que podamos formar la secuencia ganadora, en la menor cantidad de manos posibles.

Probabilidades independientes y condicionales (3)

Si nos ponemos más formales, una probabilidad condicional equivale a decir que:

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$



Recordando las matemáticas de la secundaria, esto se puede representar con un diagrama de Venn:

Hay que notar que: $P(A,B)=P(A|B) \cdot P(B)$

También: $P(A,B) = P(B,A)$

Ejercicio

Con base en la probabilidad de que una figura sea una X, esto es, P(X), ¿cuántas probabilidades hay de que dada una X sea roja, es decir, P(X | roja)?

X	X	X	O	O
X	X	O	X	O
O	O	X	O	X
O	O	O	X	O
X	X	X	X	O

Probabilidades en lingüística, *again*… (1)

Cuando vemos el lenguaje desde un punto de vista probabilístico, podemos considerar que los fenómenos que ocurren en él son de carácter condicional.

Por ejemplo, Jurafsky y Martin (2000) proponen el siguiente ejemplo: ¿cuál es el grado de probabilidad para construir una oración como *I want to eat Chinese food lunch*.

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Probabilidades en lingüística, *again*… (2)

Cardie (2006) explica que, básicamente, los n-gramas permiten hacer conteos de palabras bajo la hipótesis de que, si identificamos una palabra con la cual inicie una cadena sintáctica (frase, oración, párrafo, etc.), podemos inferir probabilísticamente cuál es la que le sigue, usando una fórmula como:

- $P(w_1, w_2, \ldots, w_{n-1}, w_n)$

$$P(w_1^n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1^2)\ \ldots P(w_n|w_1^{n-1})$$

$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

Probabilidades en lingüística, *again*… (3)

Volviendo con el ejemplo de Jurafsky y Martin (2000), y considerando entonces que, dependiendo de la palabra que hayamos identificado dentro de una cadena sintáctica, ya se puede proponer cierto grado de probabilidad:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Probabilidades en lingüística, *again*… (4)

En términos simples, lo que esperamos es que nuestros *n-gramas* se puedan representar del siguiente modo:

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>

Gracias por su atención

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