Apunte de Probabilidad y Estadística (C) 1C 2018

Versión del 28 de marzo de 2018

Profesor: Matthieu Jonckheere.

Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Entre otras Fuentes:

Notas originales de P. Ferrari.

Ana Bianco, Elena Martínez (2004), Probabilidades y Estadística (Computación)

Sheldon Ross (1997), A first course in Probability.

Contenidos

1. Espacios de probabilidad

Experimentos aleatorios y determinísticos

Descripción de un espacio de probabilidad $(S, \mathcal{F}, P)$.

a) $\mathcal{S}$ Espacio muestral:

resultados posibles de un experimento aleatorio.

Ejemplos:

Moneda: $\mathcal{S} = \{\text{Cara,Seca}\} = \{1, 0\}$

Dado: $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$

Dos monedas

10 monedas: $\mathcal{S} = \{0, 1\} \times \cdots \times \{0, 1\}$ (diez veces)

infinitas monedas: $\mathcal{S} =$ todas las sucesiones de 0 y 1.

Dos dados $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}^2$.

Tiempo de vida de una lámpara $\mathcal{S} = [0, \infty)$.

b) Eventos o sucesos $\mathcal{F}$

Subconjuntos de $\mathcal{S}$ en el caso numerable.

Ejemplos:

Cara sola, seca sola

Dos dados: suma par, suma igual a 7, resta menor que 2

10 monedas: por lo menos 5 caras.

lampara dura entre 3 y 5 meses

Operaciones con eventos

Unión, intersección, uniones e intersecciones numerables, complementos.

$\mathcal{S}$ es el evento cierto o seguro.

$\emptyset$ es el evento imposible.

$A \cup B$ Unión: Ocurre A ó B.

$A \cap B$ Intersección: Ocurre A y B.

A^c Complemento de A. No ocurre A.

$A - B = A \cap B^c$. Diferencia: Ocurre A y no ocurre B.

$A \subset B =$ si A ocurre implica que B ocurre.

$A \cap B = \emptyset$ si A y B son mutuamente excluyentes o disjuntos.

Propiedades:

Asociatividad: $A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$

$A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$

Conmutatividad: $A \cup B = B \cup A, \quad A \cap B = B \cap A$

Distributividad: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Leyes de De Morgan:

$$\left(\cup_i A_i\right)^c = \cap_i A_i^c, \qquad \left(\cap_i A_i\right)^c = \cup_i A_i^c$$

c) Probabilidad

Interpretación intuitiva de la Probabilidad: Se repite n veces un mismo experimento aleatorio en forma independiente y bajo las mismas condiciones.

n_A: número de veces que ocurre A.

Frecuencia relativa de A en n experimentos:

$$f_n(A) = \frac{n_A}{n}$$

La evidencia empírica muestra que cuando n crece, $f_n(A)$ tiende a estabilizarse alrededor de un número $P(A)$:

$$\lim_{n\to\infty} f_n(A) = P(A).$$

Propiedades

1) $f_n(A)$ está entre 0 y 1

2) $f_n(\mathcal{S}) = 1$

3) Si $A \cap B = \emptyset$,

$$f_n(A \cup B) = \frac{n_{A\cup B}}{n} = \frac{n_A}{n} + \frac{n_B}{n} = f_n(A) + f_n(B).$$

Axiomas de Probabilidad: Experimento, espacio muestral $\mathcal{S}$.

A cada evento A se le asocia $P(A)$, llamada *probabilidad de A*

$P(A)$ debe satisfacer los siguiente **axiomas**:

A1. $P(A) \in [0, 1]$ para todo evento A.

A2. $P(\mathcal{S}) = 1$

A3. Si $A_1, A_2, \ldots$ disjuntos (es decir si $A_i \cap A_j = \emptyset$, para todo $i \neq j$), entonces

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Ejemplo: Moneda. $\mathcal{S} = \{cara, ceca\} = \{1, 0\}$. $P(\{1\}) = p$ y $P(\{0\}) = 1 - p$, $P(\{0, 1\}) = 1$, $P(\emptyset) = 0$, con $0 \leq p \leq 1$, satisface los axiomas.

Propiedades:

1) $P(A^c) = 1 - P(A)$ para todo evento A

2) $P(\emptyset) = 0$

3) Si $A \subset B \Rightarrow P(A) \leq P(B)$ y $P(B - A) = P(B) - P(A)$

Dem: Si $A \subset B \Rightarrow B = A \cup (B - A)$ y éstos dos eventos son excluyentes. Por el axioma A3 $P(B) = P(A) + P(B - A)$. Como por el axioma A1, $P(B - A) \geq 0$, resulta $P(B) \geq P(A)$ y, despejando, se obtiene la segunda afirmación.

3') De la misma manera, mas generalemente, $B = (A \cap B) \cup (B - A)$ y

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

$$P(B - A) = P(B) - P(A \cap B).$$

Simetricamente:

$$P(A - B) = P(A) - P(A \cap B).$$

4) Dados dos eventos cualesquiera A y B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Dem: $A \cup B = (A - B) \cup (B - A) \cup (A \cap B)$ y estos eventos son excluyentes. Por A3:

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

y podemos concluir usando (3').

5) Dados dos eventos cualesquiera A y B, $P(A \cup B) \leq P(A) + P(B)$.

Dem: Esta propiedad se deduce inmediatamente de la propiedad anterior y del axioma A1.

Ejercicios: a) Demostrar, usando la propiedad 4) que, dados tres eventos cualesquiera,

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$

$$-P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_1 \cap A_2) + P(A_1 \cap A_2 \cap A_3)$$

b) Probar, usando inducción que, dados $A_1, A_2, \dots$ eventos cualesquiera,

$$P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

Asignación de probabilidades: Si el espacio muestral $\mathcal{S}$ es finito o infinito numerable, llamamos *eventos elementales* a los conjuntos $\{x\}$ con $x \in \mathcal{S}$. Así $\mathcal{S} = \cup_{x \in \mathcal{S}}^{\infty} \{x\}$.

Si conocemos $p_x = P(\{x\})$, de manera que $\sum_{i=1}^{\infty} p_x = 1$, entonces para cualquier evento $A = \cup_{x \in A}^{\infty} \{x\}$,

$$P(A) = \sum_{x \in A}^{\infty} p_x.$$

Ejemplos: 1) Dado equilibrado. $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ y $p_i = 1/6$ para $i = 1, \dots, 6$.

Para calcular $P(A) = P($ resultado par$) = P(2, 4, 6)$, se obtiene $P(A) = p_2 + p_4 + p_6 = 1/2$

2) Construya una probabilidad en el espacio $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ tal que la probabilidad de las caras pares es el doble que la probabilidad de las caras impares:

$p_1 + p_3 + p_5 = p, \qquad p_2 + p_4 + p_6 = 2p$

Como $P(\mathcal{S}) = 1$, $3p + 3\,2p = 1$, entonces $p = 1/9$.

3) Arrojamos una moneda equilibrada 10 veces. Cual es la probabilidad que salgan exactamente 5 caras?

4) Arrojamos una moneda equilibrada hasta obtener cara. Cuál es la probabilidad de que la cara sea obtenida en un número par de lanzamientos? El espacio muestral es

$$\mathcal{S} = \{(1), (01), (001), (0001), \dots\} = \{1, 2, 3, \dots\} \tag{1}$$

y le asignamos probabilidad $p_i = \frac{1}{2^i}$, $i \geq 1$ entero. (es una probabilidad?)

El evento es $A = \{2, 4, 6, \dots\}$.

$$P(A) = \sum_{i \geq 1} p_{2i} = \sum_{i \geq 1} 1/2^{2i} = \frac{1}{1 - \frac{1}{4}} - 1 = \frac{1}{3}.$$

Espacios de equiprobabilidad: $\mathcal{S}$ es finito y sea $n = \#\mathcal{S}$ (el símbolo $\#$ representa el cardinal del conjunto).

Diremos que el espacio es de equiprobabilidad si los n eventos elementales tienen igual probabilidad, es decir si $p_x = 1/n$, para todo $x \in \mathcal{S}$.

Ejemplos: 1) Urna contiene 5 bolillas numeradas de 1 a 5. Retiramos dos bolillas con reposición.

Se trata de un espacio de equiprobabilidad, $\mathcal{S} = \{1,2,3,4,5\} \times \{1,2,3,4,5\}$ entonces su cardinal es $\#\mathcal{S} = 5 \times 5 = 25$.

Supongamos que las bolillas 1 y 2 son blancas y las otras 3 rojas.

a) ¿Cuál es la probabilidad de que se extraiga al menos una bolilla roja?

b) ¿Cuál es la probabilidad de que la primera bolilla extraída sea roja y la segunda blanca?

El evento ninguna roja es $A^c = \{12, 21, 11, 22\}$ tiene 4 elementos. Así $P(A) = 1 - P(A^c) = 21/25$.

b) A tiene 3×2 elementos. Así $P(A) = 6/25$.

Observe que el espacio "color de las dos bolillas ordenado" $\{BB, BR, RB, RR\}$ no es equiprobable en este caso.

2) Lanzamiento de n monedas. Sucesiones de n ceros y unos.

Si la moneda es honesta $\mathcal{S}$ tiene 2^n elementos y cada uno tiene probabilidad $1/2$.

Cual es la probabilidad que la primera moneda sea cara?

Cual es la probabilidad que la quinta moneda sea cara?

Cual es la probabilidad que la séptima sea cara y la novena sea ceca?

3) Problema de las 3 puertas. Tres puertas cerradas y un premio atras de una de las puertas. Elijo una puerta y el presentador abre una de las otras dos que no tiene premio. Me da la opcion de cambiar de puerta. Conviene cambiar?

4) Una pequeña comunidad tiene 10 madres, cada una con 3 hijos. Una madre y uno de sus hijos van a ser elegidos al azar. Cual es la probabilidad que la madre más joven y su hijo mayor sean elegidos?

2. Probabilidad condicional e independencia

100 personas

2 enfermos y no vacunados
4 enfermos y vacunados
76 sanos y vacunados
18 sanos y no vacunados

Elijo una persona al azar de esa población y observo su estado.

El espacio muestral es $\mathcal{S} = \{ev, en, sv, sn)$ y los eventos son

$E = \{ev, en)$ (enfermo),

$V = \{ev, sv)$ (vacunado).

$P(\{ev\}) = 0{,}04$, $P(\{en\}) = 0{,}02$, $P(\{sv\}) = 0{,}76$, $P(\{sn\}) = 0{,}18$ (cálculos hechos con casos favorables sobre posibles)

Cual es la probabilidad que una persona esté enferma?

$P(E) = P(\{ev, en\}) = 0{,}04 + 0{,}02 = 0{,}06$.

Probabilidad que una persona vacunada esté enferma?

Casos favorables 4, casos posibles $76 + 4$ (los vacunados)

Si sabemos que la persona elegida está vacunada, cual es la probabilidad que esté enferma?

Hay que restringir el espacio muestral a los vacunados.

$P(\text{enfermo dado vacunado}) = \frac{4}{80} = P(EV)/P(V)$

Definición de Probabilidad condicional: $\mathcal{S}, P$, Eventos A, B con $P(B) > 0$

$P(A|B) = P(AB)/P(B)$ es la *proba condicional* de A dado que conocemos B.

Observaciones

- $P(AB) = P(A|B)P(B)$
- $(B, P(\cdot|B))$ nuevo espacio de proba.

Ejemplos

Dados

Un dado. Calcule la probabilidad de ver un 3 dado que el resultado es a lo sumo 4.

Dos dados. Calcule la probabilidad de que haya salido un seis dado que la suma es mayor o igual a 9.

Monedas Lanzamos 3 monedas. Calcule la probabilidad que la tercera moneda sea cara dado que el número de caras es 2.

Familias de dos hijos

$\mathcal{S} = \{vv, vm, mv, mm\}$, espacio equiprobable.

1) Una familia tiene dos hijos. Sabemos que el primer hijo es varón. Cual es la probabilidad que el segundo hijo sea también varón?

$A = \{vv\}$ (dos hijos varones), $C = \{vv, vm\}$ (primer hijo varón),

Queremos calcular $P(A|C) = P(AC)/P(C) = \frac{1/4}{2/4} = 1/2$

2) Sabemos que una familia conocida con dos hijos tiene por lo menos un hijo varón. Cual es la proba que los dos sean varones?

Buscamos $P(A|C)$, con $A = \{vv\}$ (dos hijos varones), y $C = \{vv, vm, mv\}$ (por lo menos un varón).

Usando las fórmulas $P(A|C) = P(AC)/P(C) = \frac{1/4}{3/4} = 1/3$.

3) Supongamos que visitamos a la familia, tocamos el timbre y un chico varón abre la puerta. Cual es la probabilidad que el otro chico sea varón?

$\mathcal{S} = \{v^*v, vv^*, m^*v, mv^*, v^*m, vm^*, m^*m, mm^*\}$

donde * quiere decir "abrió la puerta". Por ejemplo mv^* es el evento que el primer hijo es mujer, el segundo hijo es varón y es él quien abre la puerta. Espacio equiprobable.

Buscamos $P(A|C)$, donde $A = \{v^*v, vv^*\}$ (los dos hijos son varones) y $C = \{v^*v, vv^*, mv^*, v^*m\}$ (abre la puerta un varón)

$P(A|C) = \frac{P(AC)}{P(C)} = \frac{2/8}{4/8} = 1/2$.

Regla de la multiplicación Cálculo de probabilidades usando árboles

$$P(A_1 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \dots P(A_n|A_1 \dots A_{n-1})$$

Dem: Por inducción. $P(A_1A_2) = P(A_1)P(A_2|A_1)$, por definición. $P(A_1 \dots A_n) = P(A_1 \dots A_{n-1})P(A_n|A_1 \dots A_{n-1})$ (por el caso de dos conjuntos) y la prueba sale aplicando la hipótesis inductiva a $P(A_1 \dots A_{n-1})$. □

Ejemplo Una urna contiene tres bolillas con etiquetas a,b,c. Hacemos 3 extracciones con reposición.

Cual es la probabilidad que las 3 letras extraídas sean diferentes?

Defina los eventos:

A_1 = la primera bolilla retirada tiene cualquier letra.

A_2 = la letra de la segunda bolilla retirada es diferente de la primera.

A_3 = la letra de la tercera bolilla retirada es diferente de las anteriores.

El evento "las tres letras diferentes" es $A = A_1A_2A_3$.

$P(A) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) = \frac{3}{3}\frac{2}{3}\frac{1}{3}$

Fórmula de la probabilidad total

Una *partición* de $\mathcal{S}$ es una familia de conjuntos disjuntos dos a dos B_i tal que

$$\mathcal{S} = \dot{\cup}_i B_i$$

En ese caso $P(\mathcal{S}) = \sum_i P(B_i)$

Ejemplo. Dado. $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.

$B_1 = \{1, 2\}$, $B_2 = \{3, 4, 5\}$ es una partición de $\mathcal{S}$.

Teorema de la Probabilidad total *Sea B_i una partición de $\mathcal{S}$ tal que $P(B_i) > 0$ para todo i. Sea A un evento. Entonces,*

$$P(A) = \sum_i P(A|B_i)P(B_i).$$

Dem $P(A) = P(\cup_i(A \cap B_i)) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$.

Ejemplo Engripados y vacunados. 80 % de la población está vacunada. De los vacunados 2 % se enferman de gripe. De los no vacunados, 10 % se enferman.

Cual es la probabilidad que una persona tenga gripe?

E = enfermo, $P(E) = ?$

V = vacunado; V^c = no vacunado

Conocemos $P(V) = 0{,}8$, $P(E|V) = 0{,}02$, $P(E|V^c) = 0{,}10$.

Usando probabilidad total:

$$\begin{aligned} P(E) &= P(E|V)P(V) + P(E|V^c)P(V^c) \\ &= 0{,}02\ 0{,}8 + 0{,}10\ 0{,}2 = 0{,}036 \end{aligned}$$

Cual es la proba que una persona con gripe haya sido vacunada?

$$P(V|E) = \frac{P(E|V)P(V)}{P(E)} = \frac{0{,}2\ 0{,}8}{0{,}036} = 0{,}44444$$

Cual es la proba que una persona con gripe no haya sido vacunada?

$$P(V^c|E) = \frac{P(E|V^c)P(V^c)}{P(E)} = \frac{0{,}10\ 0{,}2}{0{,}036} = 0{,}5555$$

Fórmula de Bayes

Sea B_i una partición de $\mathcal{S}$ y sea A un evento con $P(A) > 0$. Entonces,

$$P(B_j|A) = \frac{P(B_jA)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Se usa cuando sabemos calcular $P(A|B_i)$ y $P(B_i)$

Juego de las 3 puertas B_i = premio en puerta i. $P(B_i) = 1/3$

Jugador elige la puerta 1 (los otros casos son análogos).

A = presentador abre la puerta 3 (el otro caso es análogo).

$P(A|B_3) = 0$, $P(A|B_2) = 1$, $P(A|B_1) = 1/2$.

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)$$
$$= \frac{1}{2}\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3} = \frac{1}{2}$$

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{1/6}{1/2} = 1/3.$$
$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A)} = \frac{1/3}{1/2} = 2/3.$$

O sea que P(No cambiar de puerta y ganar) $= 1/3$ y

P(Cambiar de puerta y ganar) $= 2/3$

Independencia

Los eventos A y B son *independientes* si $P(AB) = P(A)P(B)$

Justificación: de las fórmulas vemos que

A y B son independientes si y sólo si $P(A|B) = P(A)$.

Es decir que el conocimiento de que B ocurrió no cambia la probabilidad que A ocurra.

Ejemplos. Dos dados. A = suma 6. F = primer dado 4. No son independientes.

B = suma 7. F y B son independientes.

Ejercicios: a) Probar que si A B son independientes, entonces A y B^c también lo son.

b) bajo qué condiciones A y A^c son independientes?

c) Sea $S = \{1, \ldots, n\}$ un espacio equiprobable y $A, B \subset S$. Demuestre que si n es primo y A y B son independientes, entonces

$$\{A, B\} \cap \{\emptyset, S\} \neq \emptyset, \tag{2}$$

es decir, o A o B es vacío o todo el espacio.

Denotemos $P(A) = a/n$, $P(B) = b/n$. $P(AB) = c/n$ con $a, b, c \in \{0, 1, \ldots, n\}$. Si los eventos son independientes, tenemos que $ab = nc$, con $c \leq a$, $c \leq b$ y $a, b, c \leq n$. Eso implica que $[a = c = 0$ ó $b = c = 0]$ ó $[0 \notin \{a, b, c\}]$, pero en este caso, como n es primo, n divide a a ó b, digamos que n divide a a, pero como $n \geq a$, necesariamente $a = n$.

d) Demostrar que si el espacio no es equiprobable, es posible encontrar A y B independientes con probabilidades estrictamente entre 0 y 1 aunque n sea primo.

Proponemos $n = 5$, $p(1) = p(2) = p(3) = \frac{1}{4}$, $p(4) = p(5) = \frac{1}{8}$. Definimos $A = \{1, 2\}$ y $B = \{1, 3\}$. Entonces $P(AB) = p(1) = \frac{1}{4} = \frac{1}{2}\frac{1}{2} = P(A)P(B)$.

Familia de eventos independientes

Tres eventos A, B, C son *independientes* si

$P(ABC) = P(A)P(B)P(C)$, $P(AB) = P(A)P(B)$, $P(AC) = P(A)P(C)$, $P(CB) = P(C)P(B)$

Si A, B, C son independientes entonces A es independiente de cualquier evento formado a partir de B y C.

Por ejemplo: C es independiente de $A \cup B$:

$$P(C \cap (A \cup B)) = P(CA) + P(CB) - P(CAB)$$

$$= P(C)[P(A) + P(B) - P(AB)] = P(C)P(A \cup B).$$

Sea J un conjunto discreto de índices, es decir un conjunto finito o infinito numerable. Los eventos de una familia $(A_j : j \in J)$ son *independientes* si

$$P(\cap_{i\in K} A_i) = \prod_{i\in K} P(A_i)$$

para cualquier subconjunto **finito** de índices $K \subset J$. Observe que sólo se pide que la probabilidad de la intersección de eventos sea el producto de las probabilidades para subfamilias finitas de conjuntos!

Ejemplos: *Infinitas monedas* Este modelo se define dando probabilidades a los eventos que dependen de un numero finito de monedas:

A_i = la i-ésima moneda es cara = sucesiones de 0's y 1's que tienen un 1 en la posición i.

Se establece que $P\big((\cap_{i\in I} A_i) \cap (\cap_{j\in J} A_j^c)\big) = p^{\#I}(1-p)^{\#J}$ si $I \cap J = \emptyset$.

Por ejemplo $P(A_k) = P(\text{cara en la jugada } k) = p$.

$P(A_k A_{k+3}^c) = P(\text{cara en la jugada } k \text{ y ceca en la jugada } k+3) = p(1-p)$.

$P(A_1 \dots A_k) = p^k$.

$B_k :=$ la primera cara ocurre en el lanzamiento k.

$B_k = P(A_1^c \dots A_{k-1}^c A_k) = (1-p)^{k-1}p$

Dados. dos dados son lanzados simultaneamente hasta que la suma de sus faces sea 5 o 7. Cual es la probabilidad que cuando aparece suma igual a uno de esos dos valores, la suma de las faces sea 5? O sea, que aparezca suma 5 antes de suma 7.

E_n = no aparece ni suma 5 ni suma 7 en los primeros $n-1$ ensayos y aparece suma 5 en el n-ésimo ensayo.

Estamos calculando

$$P(\cup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n) = (*)$$

porque los eventos son disjuntos.

Sean A_j = suma 5 en la jugada j, B_j = suma 7 en la jugada j.

$H_j = (A_j \cup B_j)^c$ = no sale ni suma 5 ni suma 7 en la jugada j.

$P(A_j) = 4/36$, $P(B_j) = 6/36$, $P(A_j \cup B_j) = 10/36$, $P(H_j) = 26/36$.

Eventos dependientes de j distintos son mutuamente independientes.

$$E_n = H_1 \dots H_{n-1} A_n$$

Por independencia:

$$P(E_n) = P(H_1 \dots H_{n-1} A_n) = \Big(1 - \frac{10}{36}\Big)^{n-1} \frac{4}{36}$$

Así

$$(*) = \sum_{n=1}^{\infty} \Big(1 - \frac{10}{36}\Big)^{n-1} \frac{4}{36} = \frac{2}{5}.$$

Solución usando probabilidad condicional Condicionamos a lo que ocurre en el primer ensayo:

$$P(E) = P(E|A_1)P(A_1) + P(E|B_1)P(B_1) + P(E|H_1)P(H_1)$$

$P(E|A_1) = 1$, $P(E|B_1) = 0$, $P(E|H_1) = P(E)$. O sea:

$$P(E) = 1\,P(A_1) + 0\,P(B_1) + P(E)P(H_1)$$

de donde

$$P(E) = \frac{P(A_1)}{1 - P(H_1)} = \frac{P(A_1)}{P(A_1) + P(B_1)} = \frac{2}{5}$$

Eventos independientes dos a dos pero no independientes.

2 monedas

A_1 primera moneda cara.

A_2 segunda moneda cara.

A_3 las dos monedas son iguales.

Son independientes dos a dos pero no independientes.

3. Variables aleatorias, acumulada, esperanza, varianza

Una variable aleatoria es una función $X : \mathcal{S} \to \mathbb{R}$. Sirve para describir eventos en $\mathcal{S}$.

Notación $\{X \in A\} = \{s \in \mathcal{S} : X(s) \in A\}$

Variable aleatoria **discreta** asume numerables valores con probabilidad positiva.

$R(X) :=$ Rango de $X = \{x \in \mathbb{R} : P(X = x) > 0\}$.

Induce una partición de $\mathcal{S}$:

$$\mathcal{S} = \dot{\cup}_{x \in R(X)} \{s \in \mathcal{S} : X(s) = x\}$$

Función de probabilidad puntual $p_X(x) = P(X = x)$ (o **distribución**)

Es una tabla.

Ejemplo Dos monedas, $\mathcal{S} = \{00, 01, 10, 11\}$. $X =$ número de caras. $X(00) = 0$, $X(01) = X(10) = 1$, $X(11) = 2$.

Induce la partición: $\{X = 0\} = \{00\}$, $\{X = 1\} = \{01, 10\}$, $\{X = 2\} = \{11\}$

Permite calcular la distribución:

x	0	1	2
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Ejemplos *Suma de dos dados.*

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Geométrica de parámetro p.

$P(X = x) = (1 - p)^{x-1} p$ con x entero positivo.

Tiro al blanco unidimensional discreto. Se elige un punto al azar en el conjunto $\mathcal{S} = \{-3, -2, -1, 0, 1, 2, 3\}$. Considere la variable aleatoria X que a cada punto $w \in \mathcal{S}$ le asigna su distancia al origen. Tenemos entonces que X toma valores en $\{0, 1, 2, 3\}$ con probabilidades $1/7$ para el origen y $2/7$ para los otros valores.

Diagrama de barras: gráfico de la función $x \mapsto P(X = x)$.

Histograma: A cada x del rango se le asigna un rectángulo cuyo área es igual a $P(X = x)$. Hay cierta libertad sobre cómo asignar la base del rectángulo.

Función de distribución acumulada

Def. $F_X(x) := P(X \leq x)$

Propiedades de la función de distribución acumulada: $F = F_X$

i) para todo $x \in \mathbb{R}$, $F(x) \in [0,1]$

ii) F es monótona no decreciente: $x \leq y$ implica $F(x) \leq F(y)$

iii) F es continua a derecha, es decir $\lim_{h\to 0^+} F(x+h) = F(x)$

iv) $\lim_{x\to\infty} F(x) = 1$ y $\lim_{x\to-\infty} F(x) = 0$

v) Altura del salto = probabilidad puntual: $p(x) = F(x) - F(x-)$

donde $F(x-) = \lim_{h\to 0} F(x-h)$

Uso La distribución acumulada de X caracteriza la función de probabilidad puntual. de X

$$P(a < X \leq b) = F(b) - F(a)$$

$$P(a \leq X \leq b) = F(b) - F(a-)$$

$$P(a \leq X < b) = F(b-) - F(a)$$

$$P(a < X < b) = F(b-) - F(a-)$$

Ejemplo. Distribución geométrica de parámetro p

Ensayos de Bernoulli con probabilidad $p \in (0,1)$ de éxito.

Exito con proba p, fracaso con proba $1-p$.

Definimos X como el número de experimentos hasta el primer éxito.

La probabilidad puntual está dada por:

$p_X(k) = P(X = k) = (1-p)^{k-1}p$. (Ya sabemos que la suma es 1.)

$P(X > k) =$ proba de k fracasos en las primeras k jugadas $= (1-p)^k$.

Así $F(k) = P(X \leq k) = 1 - P(X > k) = 1 - (1-p)^k$

Graficar la función de probabilidad puntual y la acumulada con $p = 1/2$.

Mostrar que los saltos son las probas.

Esperanza La esperanza de una variable aleatoria es definida como

$$EX = \sum_x xP(X = x)$$

(si la suma con el módulo existe $\sum_x |x|P(X = x) < \infty$)

La suma es sobre el rango $R_X = \{x : P(X = x) > 0\}$

Ejemplos: 1) $X =$ dado; $EX = 3{,}5$.

2) número de caras en 2 monedas. $EX = 1$.

3) variable Bernoulli(p). $EX = P(X = 1) = p$

4) No existe: $P(X = x) = \frac{6}{\pi^2}\frac{1}{x^2}$.

Interpretaciones

Centro de gravedad.

Ley de grandes números.

Opciones ante un evento aleatorio

Billete de lotería vale \$1 con premio $\$10^6$.

Probabilidad de ganar es $1/10^7$ (hay 10 millones de billetes).

$\mathcal{S} = \{0, 1\}$, donde 1 = gana el billete, 0 = pierde el billete.

$$P(\{1\}) = \frac{1}{10^7}, \qquad P(\{0\}) = 1 - \frac{1}{10^7}$$

Opción 1: comprar el billete; lucro $X(1) = 10^6$–1, $X(0) = -1$

$$EX = \frac{1}{10^7}(10^6\text{–}1) + (1\text{–}\frac{1}{10^7})(-1) = -0{,}9$$

Opción 2: No comprar el billete: lucro $Y(1) = Y(0) = 0$

$$EY = 1(0) = 0,$$

"No podés perder si no jugás".

Mintiendo con estadística

Un colegio tiene 3 aulas, con 5, 10 y 150 alumnos, respectivamente.

X = número de alumnos de un aula elegida al azar

$\mathcal{S} = \{1, 2, 3\}$ equiprobable: $X(1) = 5$, $X(2) = 10$, $X(3) = 150$.

Cual es el tamaño promedio del aula

$$EX = \frac{1}{3}5 + \frac{1}{3}10 + \frac{1}{3}150 = \frac{165}{3} = 55$$

Número promedio de estudiantes por aula es 55.

Ahora elija *un estudiante* y vea de que tamaño es su aula.

$$\mathcal{S} = \{1, 2, \dots, 165\}, \qquad \text{equiprobable}$$

Y = tamaño del aula de un estudiante elegido al azar.

$$Y(k) = \begin{cases} 5, & \text{si} \quad k \leq 5 \\ 10, & \text{si} \quad 11 \leq k \leq 20 \\ 150, & \text{si} \quad 21 \leq k \leq 165 \end{cases}$$

$P(Y = 5) = \frac{5}{165}$, $P(Y = 10) = \frac{10}{165}$, $P(Y = 150) = \frac{150}{165}$.

$$EY = \frac{5}{165}5 + \frac{10}{165}10 + \frac{150}{165}165 = 137$$

es el tamaño promedio del aula del estudiante elegido al azar.

Esperanza de la geométrica(p)**:** $P(X = k) = (1-p)^{k-1}p$, $k = 1, 2, \dots$

$$EX = \sum_{k\geq 1} k(1-p)^{k-1}p = -p\sum_{k\geq 1}((1-p)^k)' = -p\Big(\sum_{k\geq 1}(1-p)^k\Big)'$$

$$= -p\Big(\frac{1}{1-(1-p)} - 1\Big)' = -p\Big(\frac{1}{p} - 1\Big) = -p\Big(-\frac{1}{p^2}\Big) = \frac{1}{p}$$

Lema Si $X \geq 0$, entonces

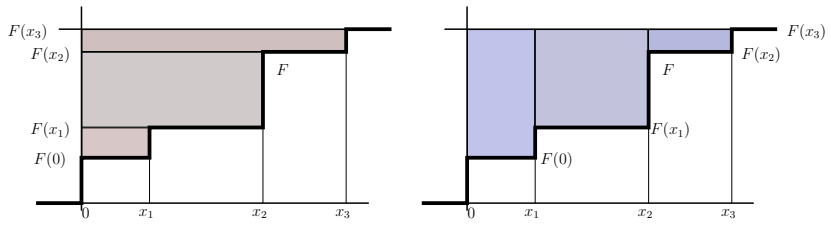
$$EX = \int_0^\infty (1 - F(x))dx$$

Demostración. Sea X discreta con rango $R_X = \{x_0, x_1, x_2, \dots\}$ con $x_0 = 0$ y $x_{i-1} < x_i$ para todo $i \geq 1$.

Como $p(x_i) = F(x_i) - F(x_{i-1})$, tenemos

$$\begin{aligned} EX &= \sum_{i=1}^{\infty} x_i p(x_i) = \sum_{i=1}^{\infty} x_i (F(x_i) - F(x_{i-1})) \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{i} (x_j - x_{j-1})(F(x_i) - F(x_{i-1})) \\ &= \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} (x_j - x_{j-1})(F(x_i) - F(x_{i-1})) \qquad \text{(Fubini)} \\ &= \sum_{j=1}^{\infty} (x_j - x_{j-1})(1 - F(x_{j-1})) = \int_0^{\infty} (1 - F(x))dx \end{aligned}$$

Esto es una integración por partes discreta. Se llama sumas de Abel. Si 0 no está en el rango, tendremos $F(0) = 0$ y exactamente la misma cuenta vale. □



Las figuras muestran dos maneras de expresar $\int_0^{\infty}(1 - F(x))dx$ en el caso discreto. A la izquierda $\sum_{i\geq 1} x_i(F(x_i) - F(x_{i-1}))$ y a la derecha $\sum_{j\geq 1}(x_j - x_{j-1})(1 - F(x_{j-1}))$.

Para la geométrica, como $x_i = i$, vale

$$EX = \sum_{x\geq 0} P(X > x) = \sum_{x\geq 0} (1-p)^k = \frac{1}{1-(1-p)} = \frac{1}{p}$$

Demostración directa de

$$EX = \sum_{x\geq 0} P(X > x)$$

$$\sum_{x\geq 0} \sum_{y\geq x+1} P(X = y) = \sum_{y\geq 1} \sum_{0\leq x\leq y-1} P(X = y) = \sum_{y\geq 1} yP(X = y) = EX$$

Esperanza de una función de una v.a. $Y = g(X)$

$$EY = \sum_x g(x)P(X = x)$$

Dem: Como $\{Y = y\} = \{g(X) = y\} = \dot{\cup}_{x:g(x)=y}\{X = x\}$,

$$P(Y = y) = \sum_{x:g(x)=y} P(X = x).$$

Entonces

$$\begin{aligned} EY &= \sum_y yP(Y = y) = \sum y \sum_{x:g(x)=y} P(X = x) \\ &= \sum_y \sum_{x:g(x)=y} yP(X = x) = \sum_y \sum_{x:g(x)=y} g(x)P(X = x) \\ &= \sum_x g(x)P(X = x) \end{aligned}$$

Ejemplo:

Sea X una variable uniforme en $R_X\{-2, \ldots, 4\}$. $P(X = x) = 1/7$.

Supongamos que al salir un número, me dan como premio el cuadrado de ese número. Cual es el valor esperado del premio?

El *premio* al salir x está dado por $g(x) = x^2$.

Calculo

$$\begin{aligned} Eg(X) &= \sum_{x=-2}^{4} x^2 P(X = x) \\ &= \frac{1}{7}((-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 + 4^2) \\ &= \frac{35}{7} = 5. \end{aligned}$$

Propiedades de la esperanza

1) (Linealidad) Si a y b son constantes reales, $E(aX + b) = aE(X) + b$.

Dem: Sea $h(X) = aX + b$, entonces

$$E(h(X)) = \sum_x h(x)P(X = x) = \sum_x (ax + b)P(X = x)$$

$$= \sum_x axP(X = x) + b\sum_x P(X = x) = aEX + b$$

2) Si X es una v.a. tal que $P(X = c) = 1$, entonces $E(X) = c$.

Dem: $EX = cP(X = c) = c$.

Si en el ejemplo anterior el lucro es el premio menos la apuesta, cual es el valor de la apuesta para que el lucro medio sea cero?

Si $h(x) =$ lucro cuando sale x, tenemos $h(x) = g(x) - a$. donde a es el valor de la apuesta.

El lucro medio es $Eh(X) = E(g(X) - a) = 5 - a$. Por lo tanto la apuesta debe ser igual al premio medio.

Varianza

Consideremos las siguientes distribuciones:

x	-1	0	1
P(X=x)	1/3	1/3	1/3

x	-10	0	10
P(Y=y)	1/3	1/3	1/3

z	-100	0	100
P(Z=z)	1/3	1/3	1/3

Vea que $EX = EY = EZ = 0$.

Sin embargo sus histogramas están dispersos alrededor de la media de forma diferente.

Definición 1. *La **varianza** de una v.a. X es definida por*

$$VX = E(X - EX)^2 = \sum_x (x - EX)^2 P(X = x) = \sigma^2$$

*El **desvío standard** $\sigma := \sqrt{VX}$*

Fórmula alternativa

$$VX = E(X^2) - (EX)^2$$

Dem:

La media minimiza el desvio cuadrático medio Sea X una va discreta con distribución $p(x) = P(X = x)$.

Buscamos m tal que

$$\sum_x (x-m)^2 p(x) = \text{mín}$$

Para eso derivamos en m:

$$-2\sum_x (x-m)p(x) = 0$$

De donde

$$m = \sum_x xp(x) = EX$$

Y la varianza es el valor del desvío cuadrático mínimo.

Ejemplos: 1) varianza de X Y Z arriba:

$$VX = \qquad VY = \qquad VZ =$$

2) X = número de caras pares de dos dados equilibrados

x	0	1	2
P(X=x)	1/4	1/2	1/4

3) Bernoulli.

4) Geométrica. $EX^2 - (EX)^2 = \frac{1-p}{p^2}$.

Veamos. Si X es positiva, La función de distribución acumulada de X^2 da un salto de $P(X = k)$ en el punto k^2. Reordenando las sumas obtenemos

$$\begin{aligned} EX^2 &= \int_0^\infty P(X^2 > x)dx \\ &= \sum_{k\geq 0}[(k+1)^2 - k^2]P(X > k) \\ &= \sum_{k\geq 0}[2k+1]P(X > k) \end{aligned}$$

Para la geométrica, $P(X > k) = (1-p)^k$ y obtenemos en este caso

$$\begin{aligned} EX^2 &= \frac{2(1-p)}{p}\sum_{k\geq 1} kp(1-p)^{k-1} + \sum_{k\geq 0}(1-p)^k \\ &= \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2} \end{aligned} \tag{3}$$

Y como $EX = 1/p$,

$$VX = EX^2 - (EX)^2 = \frac{2-p-1}{p^2} = \frac{1-p}{p^2}. \tag{4}$$

Propiedades de la Varianza

$$V(aX+b) = a^2 VX$$

usar formula del estadistico inconciente

Desvio standard

$$DX = \sqrt{VX}$$

$$D(aX+b) = |a|\, DX$$

Si X es constante, entonces $VX = 0$.

4. Distribuciones discretas usuales

Distribución Bernoulli y binomial Jacob Bernoulli (1654-1705), matemático suizo. Demuestra la ley débil de grandes números para variables Bernoulli.

Variable aleatoria Bernoulli: $X \in \{0, 1\}$

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

$X \sim$Bernoulli(p).

$EX = p$, $VX = p(1 - p)$

En un casino se juega al rojo representado por 1 o negro representado por 0. Cual es la probabilidad de ganar apostando al rojo en una **ruleta** con cero? $p = 18/37$.

Distribución Binomial:

El *Experimento binomial* consiste de n ensayos de Bernoulli.

Se trata de pruebas idénticas con dos resultados posibles: Éxito (1) y Fracaso (0).

Pruebas independientes.

La probabilidad de Éxito en cada prueba es constante igual a p.

Espacio muestral = {vectores de n 0's y 1's}. Un estado típico

$a = (a_1, \ldots, a_n)$, $a_i \in \{0, 1\}$.

$P(\{a\}) = P(\{(a_1, \ldots, a_n)\}) = p^{(\#1 \text{ en } a)}(1 - p)^{(\#0 \text{ en } a)}$

$S_n(a) = a_1 + \cdots + a_n$ número de éxitos en n ensayos.

$$P(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, \ldots, n$$

Dem: como la probabilidad de cada punto muestral depende solamente del número de unos,

$$P(S_n = k) = \sum_{a: S_n(a) = k} P(\{(a_1, \ldots, a_n)\})$$

$$= \sum_{a: S_n(a) = k} p^k (1 - p)^{n-k} = \#\{a : S_n(a) = k\} p^k (1 - p)^{n-k}$$

$$= \binom{n}{k} p^k (1 - p)^{n-k}$$

porque $\binom{n}{k}$ es el número de subconjuntos distintos de k objetos que se pueden elegir de un conjunto de n objetos distintos.

Veamos que la suma es uno:

$$\sum_{k=0}^{n} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1.$$

Ejemplos: 3 monedas. Cual es la probabilidad que salgan exactamente 2 caras?

5 Dados: cual es la probabilidad que salgan exactamente dos veces números menores que 3.

Defectos. Sabemos que una máquina produce piezas defectuosas con probabilidad 0.01. Cual es la probabilidad que en 100 piezas haya más de una defectuosa? Que tiene que asumir?

Motores: Suponga que un motor de avión falla con probabilidad $1 - p$ y que motores distintos fallan independientemente. Un avión vuela si por lo menos la mitad de sus motores está en funcionamiento.

Para cuales valores de p es preferible un avión de 2 motores a uno de 4 motores?

5) ¿Es el que sigue un experimento Binomial? 2 bolillas sin reposición de urna con 5 blancas y 3 negras. Éxito: "la bolilla extraída es blanca". NOOOO

Cálculo de la esperanza de la Binomial:

$$ES = \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np$$

La varianza de la Binomial es:

$$VS = np(1-p)$$

Hay que calcular $E(S(S-1))$ y de ahí sale:

$$VS = ES^2 - (ES)^2$$

Es más fácil calcular $E(S(S-1))$ y como $E(S(S-1)) = ES^2 - ES$, de ahí vamos a deducir ES^2. Veamos:

$$\begin{aligned}
&E(S(S-1)) \\
&= \sum_{k=0}^{n} \binom{n}{k} k(k-1) p^k (1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{n!}{k!(n-k)!} k(k-1) p^k (1-p)^{n-k} \\
&= n(n-1)p^2 \sum_{k=0}^{n} \frac{(n-2)!}{(k-2)!((n-2)-(k-2))!} p^{k-2} (1-p)^{(n-2)-(k-2)} \\
&= n(n-1)p^2 \sum_{k=0}^{n-2} \frac{(n-2)!}{k!((n-2)-k)!} p^k (1-p)^{n-2-k} = n(n-1)p^2
\end{aligned}$$

De donde

$$VS = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$$

Aproximación Poisson de la binomial

S_n Binomial$(n, p(n))$

$p(n) = \lambda/n$, $\lambda > 0$ es un parametro.

Lemma *Vale*

$$\lim_{n\to\infty} P(S_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Dem:

$$\begin{aligned}
P(S_n = k) = \binom{n}{k} p(n)^k (1-p(n))^{n-k} &= \frac{n!}{k!(n-k)!} \Big(\frac{\lambda}{n}\Big)^k \Big(1 - \frac{\lambda}{n}\Big)^{n-k} \\
&= \frac{\lambda^k}{k!} \Big(1 - \frac{\lambda}{n}\Big)^n \frac{n!}{(n-k)!\, n^k} \Big(1 - \frac{\lambda}{n}\Big)^{-k}
\end{aligned}$$

Pero

$$\lim_{n\to\infty} \Big(1 - \frac{\lambda}{n}\Big)^n = e^{-\lambda}$$

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\, n^k} = \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{n^k} = 1$$

$$\lim_{n\to\infty} \Big(1 - \frac{\lambda}{n}\Big)^{-k} = 1$$

Lo que prueba el Lema. □

A grandes razgos, vale para $n \geq 100$ y $p < 0,01$, lo que se dice np "moderado".

Distribución de Poisson

Simeon-Denis Poisson (1781-1840).

$\lambda > 0$ número real. X tiene distribución Poisson de parámetro λ si

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k \geq 0$$

Recordemos que por Taylor:

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Esto implica que $\sum_{k\geq 0} P(X = k) = 1$.

Cálculo de EX, VX.

$$EX = \sum_{k\geq 0} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda \sum_{k\geq 1} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda \tag{5}$$

$$E(X(X-1)) = \sum_{k\geq 0} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = \lambda^2 \sum_{k\geq 1} \frac{e^{-\lambda}\lambda^{k-2}}{(k-2)!} = \lambda^2 \tag{6}$$

De donde $EX^2 = \lambda^2 + \lambda$. Por eso,

$$VX = EX^2 - (EX)^2 = \lambda. \tag{7}$$

En otras palabras, cuando n es grande y p es chico, la distribución binomial (n, p) aproxima la Poisson(λ) con $\lambda = np$.

Ejemplos:

1. Número de errores por página de un libro. 1 cada 30 páginas.
2. Número de personas de una comunidad que llega a los 100 años. Se calcula 1 por mil.
3. Número de llamadas equivocadas que recibo en mi teléfono. 1 cada 10 dias.
4. Número de personas que van a un cajero de 12 a 12:05. 1 por minuto.

Ejemplo: si el número de errores por página de un libro es Poisson con parámetro $\lambda = 1/2$, cual es la probabilidad que una página tenga por lo menos dos errores?

Defectos. Sabemos que una máquina produce piezas defectuosas con probabilidad 0.01. Cual es la probabilidad que en 100 piezas haya más de una defectuosa? Vimos que era binomial $(100, 0{,}01)$. Aproximamos con Poisson(1).

Cálculo de la esperanza y varianza de la Poisson (λ).

$EX = \lambda$. $VX = \lambda$

La distribución de Poisson también funciona para aproximar el número de éxitos en ensayos no independientes.

Sombreros. n personas tienen n sombreros. Los sombreros se distribuyen aleatoriamente entre las n personas. Cual es la proba que el número de personas que se puso su mismo sombrero sea mayor o igual a 2?

X_n = número de coincidencias en una permutación aleatoria. La proba de éxito en cada ensayo es $1/n$, así que el número medio de éxitos es $n\,1/n = 1$. Se puede probar (ejercicio) que la distribución de X_n aproxima Poisson(1).

Binomial negativa o Pascal: Dos parametros, k y p

$$P(Y_k = t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}$$

$Y_k :=$ número de ensayos Bernoulli hasta el k-ésimo éxito.

$$EY_k = \frac{k}{p}, \qquad VY_k = \frac{k(1-p)}{p^2}$$

En ensayos independientes de Bernoulli con probabilidad p de éxito, cual es la probabilidad que ocurran por lo menos r éxitos antes de la m-ésima falla?

r éxitos ocurren antes de la m-ésima falla si el r-ésimo éxito ocurre antes del $(r+m-1)$ ensayo.

Por lo tanto la probabilidad que buscamos es

$$\sum_{n=r}^{n+m-1} \binom{n-1}{r-1} p^n (1-p)^{n-r}$$

Problema de las cajas de fósforos de Banach.

Un matemático que fuma en pipa tiene una caja con N fósforos en el bolsillo izquierdo y otra con N fósforos en el bolsillo derecho. Cada vez que prende la pipa, elige uno de los dos bolsillos al azar y usa un fósforo de la caja correspondiente. En algún momento encuentra una caja vacía, cual es la probabilidad que haya exactamente k fósforos en la otra caja? Sea A ese evento.

Solución: Sea E el evento que el matemático encuentra la caja derecha vacía y que hay k fósforos en la caja izquierda. En ese evento el matemático eligió $N+1$ veces la caja derecha y $N-k$ veces la caja izquierda, siendo que la última vez, eligió la caja derecha. Esto es exactamente $N-k$ fracasos en el instante del $N+1$ éxito o, equivalentemente $2N-k$ ensayos en el instante del $N+1$ éxito. Esto es una Binomial negativa con parámetros $N+1$ y $1/2$:

$$P(D) = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k+1}$$

Con el mismo argumento encontramos que si I es el evento que encuentra la caja izquierda vacía y k fósforos en la derecha, $P(I) = P(D)$. Como $A = D \dot{\cup} I$, tenemos

$$P(A) = P(D) + P(I) = 2P(D) = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k}$$

La geométrica no tiene memoria X geometrica(p). Entonces

$$P(X > k+n | X > k) = \frac{(1-p)^{k+n}}{(1-p)^k} = (1-p)^n = P(X > n)$$

O sea que para un colectivo que llega al cabo de un tiempo geométrico, si lo tuvimos que esperar k minutos, la probabilidad de esperarlo n minutos más es la misma que la de haber esperado más de n minutos desde el instante que llegamos a la parada.

Vimos que $EX = \frac{1}{p}$.

Cual es $E(X|X > k)$?

$$\sum_{j \geq 0} P(X > j | X > k) = k + \sum_{n \geq 0}^{\infty} P(X > n | X > k) = k + EX = k + \frac{1}{p}$$

Si esperé k minutos, en media esperaré $EX = 1/p$ minutos más (lo mismo que tenía en media cuando llegué a la parada)

5. Variables aleatorias continuas

Ejemplo: X_n: duración de una batería en unidades $1/n$.

$X_n \sim$ Uniforme en $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$.

Cuando n es grande X_n aproxima una variable aleatoria X "esencialmente continua" ("tiempo"), $X \in [0,1]$.

Histogramas con área total igual a 1.

dias, horas, minutos, segundos, décimas de segundo, etc, como límite de los histogramas una curva suave.

Probabilidad de que la duración esté entre a y b ($a < b$) estará dada por el área bajo la curva entre a y b.

$$P(X_n \in [a,b]) = \lfloor (b-a)n \rfloor \frac{1}{n} \to_{n\to\infty} b - a$$

donde $\lfloor x \rfloor$ es la parte entera de x.

Definición: Una v.a. X es **continua** si existe una función $f : \mathbb{R} \to \mathbb{R}^+ = [0, \infty)$ llamada *función de densidad* de X tal que

$$P(X \in A) = \int_A f(x)dx, \qquad A \subset \mathbb{R}$$

A Boreliano medible, etc.

Para $A = [a,b]$ (intervalo)

$$P(a \le X \le b) = \int_a^b f(x)dx$$

La función de densidad $f(x)$ debe satisfacer

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$f(x)$ puede ser mayor que 1.

Ejemplo: $f(x) = ax^2 \mathbf{1}\{x \in [1,3]\}$.

Calcular $a = \left(\int_1^3 x^2 \right)^{-1} = \frac{3}{26}$.

Calcular $P(X \ge 2) = \frac{19}{26}$

Función de distribución acumulada

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx$$

Calcular la F de la variable X

Propiedades de la función de distribución acumulada:

X v.a. continua,

i) para todo $x \in \mathbb{R}$, $F(x) \in [0,1]$.

ii) $F(x)$ es monótona no decreciente, es decir ...

iii) $F(x)$ es continua en todo punto.

iv) $\lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to\infty} F(x) = 1$

Lema 2. *Si X es continua y $a \le b$ reales, vale*

$$\begin{aligned} P(a < X < b) &= P(a \le X < b) = P(a < X \le b) \\ &= P(a \le X \le b) = F(b) - F(a) \end{aligned}$$

Demostración. Basta ver que $P(X = a) = P(X = b) = 0$. □

Lema 3. *Si X continua con densidad $f(x)$ y acumulada $F(x)$, entonces en todo punto donde $F(x)$ es derivable,*

$$f(x) = F'(x)$$

Demostración. Resulta del Teorema Fundamental del Cálculo Integral, y de la definición de $F(x)$. □

Distribución Uniforme: X tiene distribución uniforme en el intervalo $[A, B]$, si su función de densidad es

$$f(x) = \frac{1}{B-A}\mathbf{1}\{x \in [A, B]\}$$

Notación: $X \sim U(A, B)$.

Distribución acumulada está dada por:

$$F(x) = \frac{x-A}{B-A}\mathbf{1}\{x \in [A, B]\} + \mathbf{1}\{x \geq B\}$$

Note que $f(x) = F'(x)$ para todo $x \notin \{A, B\}$.

Percentiles de una distribución continua: Sea X una v.a. continua con $f(x)$ y $F(x)$ y sea $0 < p < 1$. El percentil $(100\ p)$-ésimo de la distribución de X es el valor x_p tal que

$$P(X < x_p) = p$$

$$\int_{-\infty}^{x_p} f(x) = p$$

Ejemplos (1) $f(x) = \frac{19}{26}x^2\mathbf{1}\{x \in [1, 3]\}$

$$F(x) = \frac{x^3-1}{26}\mathbf{1}\{x \in [1, 3]\} + \mathbf{1}\{x \geq 3\}$$

Percentil $p = 0,25$. $x_p \in [1, 3]$:

$$F(x_{0,25}) = 0,25 \quad \Rightarrow \quad \frac{x^3-1}{26} = 0,25 \quad \Rightarrow \quad x_{0,25} = 1,96$$

2) $X \sim$ Uniforme(A, B). Acumulada:

$$F(x) = \frac{x-A}{B-A}\mathbf{1}\{x \in [A, B]\} + \mathbf{1}\{x \geq B\}$$

Buscamos el percentil $p = 0,5$:

$$0,5 = F(x_{0,5}) \quad \Rightarrow \quad 0,5 = \frac{x_{0,5}-A}{B-A} \quad \Rightarrow \quad x_{0,5} = \frac{A+B}{2}$$

Mediana: Es el percentil $p = 0,5$.

Esperanza de una v.a. continua:

Definición: Sea X con densidad $f(x)$, la esperanza o valor esperado de X se define como

$$EX = \int_{-\infty}^{\infty} xf(x)dx = \mu_X$$

si $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$. Si no, decimos que no existe.

Ejemplo: Sea $X \sim$ Uniforme(A,B),

$$EX = \frac{A+B}{2}$$

Lema 4. *Si X tiene densidad $f(x)$ y $h : \mathbb{R} \to \mathbb{R}$, entonces*

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx$$

si la integral del modulo es finita.

Porqué esa definición de esperanza? Sea $X \in [0, K]$ una variable aleatoria continua acotada por K entero y X_n una aproximación discreta de X definida por

$$X_n = h_n(X) = \frac{k}{n}\mathbf{1}\Big\{\frac{k}{n} \le X < \frac{k+1}{n}\Big\}, \qquad k \in \{0, \dots, nK-1\}$$

X_n asume nK valores. Note que $|X_n - X| \le \frac{1}{n}$.

$$EX_n = \sum_{k=0}^{nK-1} \frac{k}{n} P\Big(X_n = \frac{k}{n}\Big) = \sum_{k=0}^{nK-1} \frac{k}{n} P\Big(\frac{k}{n} \le X < \frac{k+1}{n}\Big)$$

$$= \sum_{k=0}^{nK-1} \frac{k}{n} \int_{\frac{k}{n}}^{\frac{k+1}{n}} f(x)dx = \sum_{k=0}^{nK-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} h_n(x)f(x)dx$$

$$= \int_0^K h_n(x)f(x)dx$$

Ahora calculemos

$$|EX_n - \int xf(x)dx| \le \int_0^K |h_n(x) - x|f(x)dx \le \frac{1}{n}\int_0^K f(x)dx = \frac{1}{n}$$

O sea, si X_n converge a X y es acotada, entonces EX_n converge a EX como fue definida con la integral.

Linealidad:

Si a y b son constantes reales,

$E(aX + b) = aE(X) + b$.

Dem: Sea $h(X) = aX + b$,

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx = \int_{-\infty}^{\infty} (ax + b)f(x)dx$$

$$= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx = aE(X) + b.$$

Ejemplo: Dos especies compiten para controlar recurso dividido en dos partes con la distribución uniforme. Sea X: proporción del recurso controlada por la especie 1. X Uniforme(0,1):

$$f(x) = \mathbf{1}\{x \in [0, 1]\}$$

"vara rota" análogo a quebrar una vara en un punto aleatorio.

Cual es la proporción promedio que controla la especie que controla la mayoría del recurso.

La mayor proporción es la variable

$$h(X) = \text{máx}(X, 1 - X) = X\mathbf{1}\{X > 1/2\} + (1 - X)\mathbf{1}\{X \le 1/2\}$$

y su esperanza es

$$Eh(X) = E(X\mathbf{1}\{X > 1/2\}) + E((1 - X)\mathbf{1}\{X \le 1/2\})$$

$$= \int_{1/2}^1 xdx + \int_0^{1/2} (1 - x)dx = 3/4$$

Fórmula para la esperanza de variables positivas Sea $X \geq 0$ una variable aleatoria con densidad f y función de distribución acumulada F. Entonces

$$EX = \int_0^\infty (1 - F(y))dy \tag{8}$$

Demostración: Escribimos $x = \int_0^x dy$ y obtenemos

$$\begin{aligned}\int_0^\infty xf(x)dx &= \int_0^\infty \int_0^x dyf(x)dx \\ &= \int_0^\infty \int_y^\infty f(x)dxdy = \int_0^\infty (1 - F(y))dy\end{aligned}$$

Intercambio de integrales por Fubini porque todo es positivo.

Si X toma valores negativos y positivos, $X = X^+ - X^-$, por lo tanto

$$EX = EX^+ - EX^- = \int_0^\infty (1 - F_{X^+}(x))dx - \int_0^\infty (1 - F_{X^-}(x))dx$$

Varianza de una v.a. continua:

Definición: Sea X una v.a. continua con esperanza μ y densidad f, la varianza de X, que se denotará $V(X)$, σ^2

$$VX = E(X - EX)^2 = \int_{-\infty}^\infty (x - \mu)^2 f(x)dx$$

El *desvío standard* es la raiz de la varianza: $+\sqrt{VX}$.

Lema 5. *Vale:* $V(X) = E(X^2) - (E(X))^2$.

Linealidad:

$$V(aX + b) = a^2 VX.$$

Ejemplos: Sea $U \sim$ Uniforme$(0, 1)$, $EU = 1/2$

$$VU = E(U^2) - (EU)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

Consideremos la distribución de $X = a + (b - a)U$. Vemos que la acumulada de X es

$$F_X(x) = P(X \leq x) = P(a + (b - a)U \leq x) = P(U \leq (x - a)/(b - a))$$

que es la acumulada de $X \sim$ Uniforme(a, b). Calculemos su varianza: $EX = (a + b)/2$ y

$$VX = V(a + (b - a)U) = (b - a)^2 VU = \frac{(b - a)^2}{12}$$

6. Distribuciones usuales

Distribución normal Se dice que X tiene distribución Normal de parámetros μ y σ^2 si su función de densidad es

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x - \mu)^2}{2\sigma^2}\Big)$$

Notación: $X \sim N(\mu, \sigma^2)$. El gráfico tiene forma de campana con eje de simetría en $x = \mu$ y puntos de inflexión en $x = \mu \pm \sigma$

Es simetrica en relacion a μ: $f(\mu + x) = f(\mu - x)$

Alcanza el maximo en $x = \mu$.

Puntos de inflexión en $\pm\sigma$.

Distribución normal standard

Def: $Z \sim N(0,1)$ si $\mu = 0$ y $\sigma^2 = 1$.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big)$$

Tabulada: $Z \sim N(0,1)$, el percentil 99 de la distribución es 2.33

Propiedades:

• Si $X \sim N(\mu, \sigma^2)$ entonces $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Demostración:

$$F_Z(z) = P(Z \le z) = P(X \le \sigma z + \mu) = F_X(\sigma z + \mu)$$
$$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} F_X(\sigma z + \mu) = f_X(\sigma z + \mu)\sigma$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(\sigma z + \mu - \mu)^2}{2\sigma^2}\Big)\sigma = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z^2}{2}\Big)$$

• Si Z normal standard y $X = \sigma Z + \mu$ entonces $X \sim N(\mu, \sigma)$.

Para probar que la Normal standard es una variable aleatoria hay que verificar que $\int f = 1$. El truco es escribir

$$\Big(\int_{-\infty}^{\infty} f(x)dx\Big)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)f(y)dxdy$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}dxdy = 1,$$

pasando por coordenadas polares.

Esperanza y varianza de la normal Se calcula primero para la distribución de la normal standard Z

$$EZ = \frac{1}{\sqrt{2\pi}} \int z e^{z^2/2} dz = 0$$

Integrando impar. Integrando por partes se obtiene también:

$$VZ = EZ^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\Big(-\frac{x^2}{2}\Big) = 1$$

Se exporta para la normal $X \sim N(\mu, \sigma)$ por la formula $X = \sigma Z + \mu$:

$$EX = \mu, \qquad VX = \sigma^2$$

Cálculo de probabilidades para la Normal

Para la Normal standard, por simetría:

$$P(Z < x) = P(Z > -x)$$

Defina $\Phi(z) = P(Z \le z)$ la acumulada de la Normal standard. Está tabulada.

$X \sim N(\mu, \sigma^2)$, $(X - \mu)/\sigma \sim N(0,1)$.

$$P(X \le a) = P\Big(\frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\Big)$$
$$= P\Big(Z \le \frac{a-\mu}{\sigma}\Big) = \Phi\Big(\frac{a-\mu}{\sigma}\Big)$$

• Si Z normal standard y $X = \sigma Z + \mu$. Entonces los percentiles satisfacen

$$\frac{x_p - \mu}{\sigma} = z_p \quad y \quad x_p = z_p\sigma + \mu$$

Ejemplos

1. $X \sim N(3, 9)$. Calcular $P(2 < X < 5)$, $P(X > 0)$ y $P(|X - 3| > 6)$

$$P(2 < X < 5) = \cdots = \Phi(\frac{2}{3}) - \big(1 - \Phi(\frac{1}{3})\big) \sim 0,3779$$

2. Las notas de su examen siguen una normal de media μ y varianza σ^2. Se estima μ y σ^2 y después se dan las notas. Nota A para quien tiene tienen nota mayor que $\mu + \sigma$, nota B entre μ y $\mu + \sigma$, nota C entre $\mu - \sigma$ y μ y nota D para aquellas menores que $\mu - \sigma$. Por ejemplo $\mu = 72$, $\sigma^2 = 100$. (A rigor, no puede haber números menores que 0 ni mayores que 100, y las notas asumen valores discretos, pero la normal aquí es usada como modelo para calcular las probabilidades de los valores discretos.)

Calcule el porcentaje de alumnos que sacará cada una de las notas.

3. (Antes de la popularización de los tests de ADN) Un experto obstetra declara en un juicio de paternidad que la gestación de un bebé tiene distribución normal con parámetros $\mu = 270$ días y $\sigma^2 = 100$. El acusado puede probar que estuvo fuera del país durante un período que comenzó 290 días antes del nacimiento y terminó 240 días antes del nacimiento. En base a esta declaración, el juez declara que el acusado no es el padre. Cual es la probabilidad que el juez se haya equivocado? Es decir, cual es la probabilidad que si el acusado fue el verdadero padre, la madre haya tenido un ciclo de gestación compatible con la ausencia del padre?

X = número de días de gestación. $X \sim N(270, 100)$. $-X$ = fecha de comienzo del embarazo contado desde el día del nacimiento. Queremos calcular la probabilidad que $-X$ sea menor que -290 o mayor que -240.

$$P(-X < -290) + P(-X > -240)$$

por simetría esto es igual a

$$= P(X > 290) + P(X < 240) = \cdots = 0,03,$$

las cuentas se hacen standarizando las variables y usando la tabla.

Variable exponencial Decimos que X tiene distribución exponencial de parámetro λ si su densidad es

$$f(x) = \lambda e^{-\lambda x}\mathbf{1}\{x \geq 0\}$$
$$F(x) = (1 - e^{-\lambda x})\mathbf{1}\{x \geq 0\}$$

Calculemos EX y VX

$$EX^n = \int_0^\infty x^n \lambda e^{-\lambda x} dx = \cdots = \frac{n}{\lambda} EX^{n-1}$$

Con $n = 1$ obtenemos

$$EX = \frac{1}{\lambda}, \qquad EX^2 = \frac{1}{\lambda} EX = \frac{2}{\lambda}^2$$

de donde

$$VX = \frac{1}{\lambda^2}$$

La exponencial no tiene memoria:

$$P(X > t + s | X > t) = P(X > s).$$

Ejemplo: Supongamos que el tiempo de respuesta de una terminal conectada en línea es una v.a. X con distribución exponencial con esperanza igual a 5 segundos.

a) Cuál es la probabilidad de que el tiempo de respuesta sea mayor de 10 segundos?

b) Cuál es la probabilidad de que el tiempo de respuesta esté entre 5 y 10 segundos?

c) Cual es la probabilidad que sabiendo que ya esperé 10 segundos, tenga que esperar todavía 5 segundos más?

La exponencial es límite de geométricas

Sea $Y_n \sim$ Geométrica(λ/n).

Entonces

$$P(Y_n/n \geq t) = P(Y_n \geq tn) = \Big(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$$

Distribución Gama Una variable aleatoria X tiene distribución Gama con parámetros $\alpha > 0$ y $\lambda > 0$ si su densidad es

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda e^{-\lambda x} (\lambda x)^{\alpha - 1} \mathbf{1}\{x \geq 0\}$$

donde $\Gamma(\alpha)$ está definida por

$$\Gamma(\alpha) := \int_0^\infty e^{-y} y^{\alpha - 1} dy$$

Integrando por partes se demuestra que

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

por lo que para α entero no negativo $\Gamma(\alpha) = (\alpha - 1)!$.

Cuando $\alpha = n$ es entero, X es el tiempo necesario para que haya n eventos, cuando el tiempo entre dos eventos es exponencial λ. Esto lo veremos después.

Relación de Gama con Poisson

Lema ***Sea*** $N(t)$ ***una variable Poisson de media*** λt***. Sea*** T_n ***una variable aleatoria con distribución acumulada***

$$F(t) = P(T_n \leq t) = P(N(t) \geq n)$$

entonces T_n ***tiene distribución Gama***(n, λ)**.**

Dem

$$F(t) = P(N(t) \geq n) = \sum_{j=n}^{\infty} \frac{e^{-\lambda t} (\lambda t)^j}{j!}$$

Diferenciando en t,

$$f(t) = F'(t) = \sum_{j=n}^{\infty} \frac{e^{-\lambda t} j (\lambda t)^{j-1} \lambda}{j!} - \sum_{j=n}^{\infty} \lambda \frac{e^{-\lambda t} (\lambda t)^j}{j!}$$

$$= \frac{\lambda e^{-\lambda t} (\lambda t)^{n-1}}{(n-1)!}$$

que es la densidad de la Gama(n, λ).

Ejercicio: Calcule EX y VX.

$$EX = \int_0^\infty x \frac{1}{\Gamma(\alpha)} \lambda e^{-\lambda x} (\lambda x)^{\alpha - 1} = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)\lambda} = \frac{\alpha}{\lambda}$$

VX queda como ejercicio.

Cambio de variable

Teorema *Sea X una v.a. con densidad $f_X(x)$ tal que $P(X \in (a, b)) = 1$. Sea $g : (a, b) \to \mathbb{R}$* ***estrictamente creciente*** *o bien* ***estrictamente decreciente*** *. Considere la nueva variable aleatoria $Y = g(X)$. Entonces*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \big(g^{-1}(y)\big)' \right| .$$

Dem (a) g estrictamente creciente. Calculamos la distribución acumulada de Y

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

pero como la función es estrictamente creciente en el intervalo (a, b), podemos invertirla:

$$= P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Para obtener f_Y derivamos F_Y y obtenemos

$$f_Y(y) = f_X(g^{-1}(y)) \left| \left(g^{-1}(y)\right)' \right|$$

el valor absoluto no agrega nada porque la derivada es positiva.

(b) g estrictamente decreciente. Como g^{-1} es decreciente,

$$P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

y derivando,

$$f_Y(y) = -f_X(g^{-1}(y)) \left(g^{-1}(y)\right)' = f_X(g^{-1}(y)) \left| \left(g^{-1}(y)\right)' \right|$$

porque la derivada de g^{-1} es negativa. □

Ejemplo $X \sim$ Uniforme $[0, 1]$ y $Y = X^2$. Entonces $f_Y(y) = f_X(\sqrt{y})\frac{1}{2}y^{-1/2}$.

Caso no invertible A veces, aunque g no sea invertible, se puede calcular la función de densidad de $Y = g(X)$. Por ejemplo, consideremos $X \sim$ Uniforme$[-3, 3]$ y $Y = X^2$. Como $X \in [-3, 3]$, $Y \in [0, 9]$. Calculemos F_Y y f_Y.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) = P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = 2F_X(\sqrt{y}) - 1 \end{aligned}$$

porque $F_X(x) = 1 - F_X(-x)$, por simetría de la f_X. Derivando,

$$f_Y(y) = f_X(\sqrt{y})/\sqrt{y} = \frac{1}{6}\sqrt{y}, \qquad y \in [0, 9].$$

(2) $Z \sim$ Normal$(0, 1)$ y $Y = Z^2$. Con el mismo razonamiento que en el caso anterior:

$$\begin{aligned} F_Y(y) &= 2F_X(\sqrt{y}) \\ f_Y(y) &= f_X(\sqrt{y})/\sqrt{y} \end{aligned}$$

Tratamiento de g no invertible en general

Ejemplo X uniforme en $(-1, 1)$.

$$f_X(x) = \frac{1}{2}\mathbf{1}\{x \in [-1, 1]\}$$

$Y = X^2$. Es decir $g(x) = x^2$ que no es invertible en $(-1, 1)$. Para poder aplicar el teo tenemos que particionar:

$$P(Y \leq y) = P(X^2 \leq y)$$

$$= P(X^2 \leq y, X \geq 0) + P(X^2 \leq y, X < 0)$$

$$= \int_{-\infty}^{\infty} f(x)\, \mathbf{1}\{x^2 \leq y, x \in G_1\}dx + \int_{-\infty}^{\infty} f(x)\, \mathbf{1}\{x^2 \leq y, x \in G_2\}dx$$

donde $G_1 : [0, 1]$ y $G_2 : [-1, 0]$ son tales que la función g restringida a G_i es invertible. Es decir $g^{-1}(y) \cap G_i$ tiene a lo sumo un punto.

$$= \int_0^1 f(x)\, \mathbf{1}\{x \leq \sqrt{y}\}dx + \int_{-1}^0 f(x)\, \mathbf{1}\{x \geq -\sqrt{y}\}dx$$

Derivamos en y y obtenemos

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y})\,\big|\frac{d}{dy}\sqrt{y}\big| + f_X(-\sqrt{y})\,\big|\frac{d}{dy}\sqrt{y}\big| \\ &= \frac{1}{2}\frac{1}{2\sqrt{y}} + \frac{1}{2}\frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}. \end{aligned}$$

Teorema general X variable aleatoria con densidad f_X.

Dada una función $g : \mathbb{R} \to \mathbb{R}$ medible, suponga que existen $G_1, \ldots, G_\ell$ partición de la imagen de X y $g_i : G_i \to \tilde{G}$ tales que

$g_i : G_i \to \tilde{G}$ invertibles y $g(x) = g_i(x)$ si $x \in G_i$.

Definimos $U := \sum_{i=1}^{\ell} g_i(X)\,\mathbf{1}_{G_i}(X)$.

Entonces U es continuo con densidad

$$f_U(u) = \sum_{i=1}^{\ell} f_X(g_i^{-1})(u)\,\big|\frac{d}{dy}g_i^{-1}(y)\big|\,\mathbf{1}_{\tilde{G}}(u)$$

Misma demostración.

$$P(U \le y) = \sum_{i=1}^{\ell} P(g(X) \le y, X \in G_i) = \sum_{i=1}^{\ell} P(g_i(X) \le y, X \in G_i)$$

$$\sum_{i=1}^{\ell} P(X \le g_i^{-1}(y), X \in G_i)$$

7. Generación de variables aleatorias

Cual es la probabilidad de ganar al solitario?

52 cartas. Hay 52! juegos posibles de solitario. Supongamos que tenemos una estrategia fija. Es decir, dada una de las permutaciones, hay una función $X \in \{0, 1\}$ donde X es 0 si la estrategia pierde y 1 si gana con esa permutación.

Cual es la proba de ganar? $p = P(X = 1)$.

Como hay que jugar cada permutación para saber si ganamos o perdemos, es imposible calcular la proporción de juegos en los que se gana.

Pero lo que se puede hacer es generar n juegos elegidos aleatoriamente entre las 52! permutaciones, determinar X para cada uno de los juegos y definir

$$\hat{p}_n = \frac{\#\text{juegos ganados}}{n}$$

Despues veremos que $\hat{p}_n$ converge a p en algún sentido.

Esto motiva el interés de *simular* variables aleatorias.

Generación de números seudo-aleatorios

Método de la congruencia Dados m, a, c y X_0,

$$X_{n+1} = (aX_n + c) \mod m, \qquad n \ge 0$$

X_{n+1} resto entero de dividir $X_n + c$ por m $(0 \le X_n \le m - 1)$.

Secuencia lineal congruente.

m es el módulo $m > 0$

a es el multiplicador $0 \leq a < m$

c es el incremento $0 \leq c < m$

X_0 es la semilla o valor inicial

Método multiplicativo secuencial: $c = 0$

Knuth: $m = 2^{64}$, $a = 6364136223846793005$, $c = 1442695040888963407$

Ventajas: rápidos de construir. Se puede repetir cualquier sucesión si se empieza con la misma semilla.

Desventajas: tiene ciclos. Los números se repiten idénticos cuando uno de los números generados es generado por segunda vez.

Ver wikipedia: Linear congruential generator

Generadores de números aleatorios "verdaderos"

Recomiendo fuertemente visitar la página www.random.org de donde saqué las observaciones que siguen.

PRNG (pseudo random number generator) son los generadores de números seudo aleatorios y TRNG (true random number generator) los generadores de números verdaderamente aleatorios.

"TRNG extract randomness from physical phenomena and introduce it into a computer. You can imagine this as a die connected to a computer, but typically people use a physical phenomenon that is easier to connect to a computer than a die is. A suitable physical phenomenon is atmospheric noise, which is quite easy to pick up with a normal radio. This is the approach used by RANDOM.ORG.

The process of generating true random numbers involves identifying little, unpredictable changes in the data. For example, HotBits uses little variations in the delay between occurrences of radioactive decay, and RANDOM.ORG uses little variations in the amplitude of atmospheric noise.

The characteristics of TRNGs are quite different from PRNGs. First, TRNGs are generally rather inefficient compared to PRNGs, taking considerably longer time to produce numbers. They are also nondeterministic, meaning that a given sequence of numbers cannot be reproduced, although the same sequence may of course occur several times by chance. TRNGs have no period."

Generación de variables aleatorias a partir de uniformes

A partir de ahora vamos a suponer que tenemos una fuente de números aleatorios independientes con distribución Uniforme$[0, 1]$ (continua). Construiremos funciones de uno o más de esos números para simular variables aleatorias con otras distribuciones.

Generación de una variable uniforme discreta.

Sea U Uniforme en $[0, 1]$.

Sea $V_n = [Un] + 1$ (parte entera)

Veamos que V_n es uniforme en $\{1, \ldots, n\}$:

$$P(V_n = k) = P([Un] + 1 = k) = P([Un] = k - 1)$$
$$= P(k - 1 \leq Un < k) = P(\frac{k-1}{n} \leq U < \frac{k}{n}) = \frac{1}{n}$$

En general, para generar una variable uniforme en $\{m, \ldots, m + n - 1\}$,

$$V_n = [Un] + m$$

Generacion de una permutación aleatoria Sea $n \geq 2$. Queremos generar una permutación aleatoria de los números $1, 2, \ldots, n$.

0. Inicialización: $k = n$, $X(i) = i$, $i = 1, \ldots, n$

1. Genere una uniforme V_k en $\{1, \ldots, k\}$

2. Intercambie los valores de $X(V_k)$ y $X(k)$.

3. Ponga $k \leftarrow k - 1$.

4. Si $k = 1$ imprima $X(1), \ldots, X(n)$. Si no, vuelva a 1.

Ejemplo: suponga que $n = 5$ y que $V_5 = 4$, $V_4 = 2$, $V_3 = 1$, $V_2 = 1$. En ese caso las sucesivas iteraciones del algoritmo serán las siguientes;

12345, 12354, 15324, 35124, 53124

Lema 6. *Los números $X(1), \ldots, X(n)$ son una permutación uniforme de $1, \ldots, n$.*

Demostración. Cada número tiene probabilidad $\frac{1}{n}$ de ser el último y por inducción ... □

Generación de variables aleatorias discretas Sea X una variable aleatoria discreta con probabilidad puntual

$$P(X = x) = p(x),$$

Sea U uniforme en $[0, 1]$. Sea $(J(x) : x \in R_X)$ una partición del intervalo $[0, 1]$ que satisfaga

$$|J(x)| = \text{longitud}(J(x)) = p(x).$$

Defina

$$X = \sum_x x \mathbf{1}\{U \in J(x)\}$$

Esto es equivalente a la doble implicación siguiente:

$$X = x \Leftrightarrow U \in J(x)$$

Lo que implica

$$P(X = x) = P(U \in J(x)) = |J(x)| = p(x).$$

Es decir que generamos una variable aleatoria con probabilidades puntuales $p(x)$.

Función inversa generalizada Una manera canónica de encontrar la partición J es usando la función inversa generalizada de la distribución acumulada.

Defina la función inversa generalizada por

$$F^{-1}(u) := \text{ínf}\{x : F(x) \geq u\}$$

Para X discreta con distribución acumulada F y x en el rango de X definimos

$$J(x) := [F(x-), F(x))$$

que es una partición por las propiedades de F. En ese caso tendremos

$$X = F^{-1}(U)$$

es equivalente a

$$X = x \Leftrightarrow U \in J(x).$$

Observación. La inversa generalizada coincide con la inversa cuando F es inversible. Por eso usamos la misma notación.

Ejemplo. Simule la variable X con distribucion

x	1	2	4
P(X=x)	1/2	1/4	1/4

Acoplamiento

En este contexto un *acoplamiento* de dos variables aleatorias X e Y es la simulación de ambas en función de un *mismo* número aleatorio.

Ejemplo: Queremos generar variables Y_ℓ Bernoulli con parámetro p_ℓ. Una manera es hacer lo siguiente:

$$Y_\ell = F_\ell^{-1}(U) = \mathbf{1}\{U > 1 - p_\ell\}$$

Las variables generadas tienen la distribución correcta:

$$P(Y_\ell = 1) = P(U > 1 - p_\ell) = p_\ell.$$

y satisfacen la siguiente propiedad de monotonía:

Si $p_1 \leq p_2$ entonces $Y_1 \leq Y_2$.

En general, si $1 - F_1(y) \leq 1 - F_2(y)$ para todo y y $Y_\ell := F^{-1}(U)$ entonces

$$Y_1 \leq Y_2.$$

Lo que nos dá una noción de orden entre variables aleatorias.

Ejercicio. Sucesiones de Bernoulli Construya un programa para generar una sucesión de n variables independientes Bernoulli con parametro $p \in [0, 1]$. La salida debe ser un vector de n ceros y unos.

Generación de variables aleatorias continuas

Método de inversión. Sea X una va continua con densidad f y acumulada F.

Supongamos F estrictamente creciente y continua. En ese caso la F es inversible.

Sea U una variable uniforme en $[0, 1]$.

Lema 7. *La variable $Y = F^{-1}(U)$ tiene la misma distribución que X.*

Demostración.

$$P(Y \leq a) = P(F^{-1}(U) \leq a) = P(U \leq F(a)) = F(a). \qquad \square$$

En general la F es monótona y no necesariamente estrictamente creciente. Pero si usamos la definicion de inversa generalizada, todavía vale la identidad

$$P(F^{-1}(U) \leq a) = P(U \leq F(a))$$

Generación de una exponencial λ

$F(x) = 1 - e^{-\lambda x}$, $x \geq 0$ es inversible en $[0, \infty)$.

$$F^{-1}(u) = \frac{-\log(1-u)}{\lambda}$$

Entonces la variable definida por

$$X = \frac{-\log(1-U)}{\lambda}$$

con U uniforme en $[0, 1]$ es exponencial.

Como $(1 - U)$ tiene la misma distribución que U, la variable

$$X = \frac{-\log(U)}{\lambda}$$

tambien tiene distribución exponencial.

El método del rechazo

Queremos generar una variable con densidad f.

Sabemos como generar una variable con densidad g

Sabemos que existe $c > 0$ tq

$$f(x) \leq cg(x) \qquad \text{para todo } x$$

Algoritmo del rechazo

1. Simule Y con densidad g y U uniforme en $[0, 1]$

2. Si $U \leq f(Y)/cg(Y)$, ponga $X = Y$ y termine.

   Si no, (rechace (Y, U) y) vaya a 1.

Lema 8. *La variable X así generada tiene densidad f.*

Idea de la Demostración. Defina

$$\begin{aligned} C &:= \{(x, y) : 0 \leq y \leq cg(x),\, x \in \mathbb{R}\} \\ B &:= \{(x, y) : 0 \leq y \leq f(x),\, x \in \mathbb{R}\} \\ B(a) &:= \{(x, y) : 0 \leq y \leq f(x),\, x \leq a\} \end{aligned}$$

$$\text{Area}(C) = c, \quad \text{Area}(B) = 1, \quad \text{Area}(B(a)) = \int_{-\infty}^{a} f(x)dx.$$

El vector

$$V := (Y, Ucg(Y)) \tag{9}$$

está distribuído uniformemente en C. Así, para cualquier $A \subset C$,

$$P(V \in A) = \frac{\text{Area}(A)}{\text{Area}(C)} \tag{10}$$

Aceptamos el punto propuesto cuando $V \in B$.

Por lo tanto la probabilidad de que $X < a$ al aceptar el punto está dada por la probabilidad que $V \in B(a)$ dado que $V \in B$.

$$\begin{aligned} P(X < a) &= P(V \in B(a) \,|\, V \in B) \\ &= \frac{\text{area}(B(a))/\text{area}(C)}{\text{area}(B)/\text{area}(C)} = \int_{-\infty}^{a} f(x)dx \end{aligned} \tag{11}$$

Generación de una variable normal standard Z

No se puede usar el método de inversión.

Empezamos a generar $X = |Z|$, que tiene densidad

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \qquad x \geq 0$$

Considere $g(x) = e^{-x}$, $x \geq 0$. Cuenta:

$$\frac{f(x)}{g(x)} \leq \sqrt{\frac{2e}{\pi}}$$

de donde $c = \sqrt{\frac{2e}{\pi}}$ y

$$\frac{f(x)}{cg(x)} = \exp\Big(\frac{-(x-1)^2}{2}\Big)$$

El algoritmo queda:

1. Genere Y exponencial de parametro 1, U uniforme en $[0, 1]$

2. Si

$$U \leq \exp\Big(\frac{-(Y-1)^2}{2}\Big)$$

ponga $X = Y$. Si no, vaya a (1).

Ahora defina $Z = VX - (1-V)X$, con V Bernoulli(1/2).

Z es Normal$(0,1)$.

Simplificación En el paso (2) Y es aceptada si

$$U \leq \exp\Big(\frac{-(Y-1)^2}{2}\Big)$$

que es equivalente a

$$-\log U \geq \frac{-(Y-1)^2}{2}$$

como $Y_2 = -\log U$ es exponencial (1),

1. Genere Y_1, Y_2 exponenciales (1)

2. Si $Y_2 \geq \frac{-(Y_1-1)^2}{2}$ ponga $X = Y_1$. Si no, vaya a (1).

8. Vectores aleatorios

Ejemplo Lanzamiento de una moneda dos veces. El resultado es un vector (X, Y)

Dos tipos de estudiante: el que la tira dos veces: resultados posibles $(0,0), (0,1), (1,0), (1,1)$ con proba $1/4$ cada uno.

El fiaca tira una vez y repite el resultado: $(0,0),\ (1,1)$,

Cada coordenada tiene la misma proba: $P(X=0) = P(Y=0) = 1/2$

Mirando sólo X o Y no podemos diferenciar entre los dos.

Hay que mirar el resultado de todo el vector (X, Y)

Def. Un *vector aleatorio* es una función $(X_1, \ldots, X_n) : S \to \mathbb{R}^n$.

Vectores aleatorios discretos.
Función de probabilidad conjunta:

$$p(x,y) = P(X = x, Y = y)$$

El *rango* del vector $R_{X,Y} \subset R_X \times R_Y$

$$P((X,Y) \in A) = \sum_{(x,y)\in A} p(x,y)$$

La probabilidad conjunta satisface

1) $p(x,y) \geq 0$

2) $\sum_x \sum_y p(x,y) = 1$

Distribuciones marginales Dado vector (X, Y),

$$P(X = x) = \sum_y P(X = x, Y = y), \quad \text{marginal de } X$$

$$P(Y = y) = \sum_x P(X = x, Y = y), \quad \text{marginal de } Y$$

Ejemplo Sea (X, Y) vector con distribución
$p(0,0) = 0{,}4$, $p(0,1) = 0{,}2$, $p(1,0) = 0{,}1$ y $p(1,1) = 0{,}3$.

Las marginales son

$P(X = 0) = p(0,0) + p(0,1) = 0{,}6$

$P(X = 1) = p(1,0) + p(1,1) = 0{,}4$

Toda la info en una tabla:

	0	1	X
0	0.4	0.2	0.6
1	0.1	0.3	0.4
Y	0.5	0.5	1

Vectores aleatorios continuos

Def. Un vector aleatorio $X = (X_1, ..., X_d)$ es continuo con densidad conjunta $f = f_X$ si

$$P(a_i \leq X_i \leq b_i, i = 1, \ldots, d) = \int_{a_1}^{b_1} \ldots \int_{a_d}^{b_d} f(x_1, \ldots, x_d) dx_1 \ldots dx_n$$

Así, para $A \subset \mathbb{R}^n$:

$$P((X_1, \ldots, X_d) \in A) = \int_A f(x_1, \ldots, x_d) dx_1 \ldots dx_n$$

Esto vale para A donde se pueda calcular la integral. En ese caso, en teoría de la medida se dice que A es *medible*.

Distribución acumulada

La distribución acumulada de un vector continuo se define para $x = (x_1, \ldots, x_d)$ como

$$F(x) = F(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

$$= \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} f(x_1, \ldots, x_d) dx_1 \ldots dx_d$$

Lema *La distribución acumulada de un vector caracteriza la distribución del vector.*

Dem. Basta mostrar que la acumulada conjunta determina la densidad conjunta. Lo hacemos para el caso de dos dimensiones. De la definición sigue que

$$f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y). \qquad \square$$

y "a lo físico":

$$P(x \leq X \leq x + dx, y \leq Y \leq y + dy) = \int_x^{x+dx} \int_y^{y+dy} f(z, w) dz\, dw$$

$$\sim f(x, y) dx dy$$

Distribuciones marginales Sea $X = (X_1, \ldots, X_d)$ un vector continuo con densidad f_X. Entonces cada X_i es una variable continua con densidad

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{d-1}} f_X(x_1, \ldots, x_d) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d$$

f_{X_i} es la densidad *marginal* de X_i que (por la fórmula de arriba) se obtiene integrando la densidad conjunta en todas las otras variables.

Ejemplo Sea (X, Y) vector con densidad conjunta

$$f(x, y) = \frac{1}{y} e^{-y - \frac{x}{y}} \qquad x, y > 0$$

La marginal de Y está dada por

$$f_Y(y) = \int f(x,y)dx = e^{-y}$$

para todo $y > 0$. O sea que $Y \sim \exp(1)$.

Calcule $P(X < Y)$ y $P(X < a)$

$$P(X < Y) = P((X,Y) \in A) = \int_0^\infty \int_0^y f(z,w)dzdw = \cdots = \frac{1}{3}$$

$$P(X < a) = \int_0^\infty \int_0^a f(z,w)dzdw = \cdots = 1 - e^{-a}.$$

Ejemplo (X,Y) con densidad

$$f(x,y) = \frac{1}{x}\mathbf{1}\{0 < y \leq x \leq 1\}$$

La marginal de X:

$$f_X(x) = \int_0^x f(x,y)dy = \mathbf{1}\{0 < x \leq 1\}$$

Así X tiene distribución uniforme en $(0,1]$.

La densidad de Y:

$$f_Y(y) = \int_y^1 f(x,y)dx = -\log y \mathbf{1}\{0 < y \leq 1\}$$

Ejemplo Sea U una variable uniforme en $[0,1]$. Para $u \in [0,1]$ defina $h(u) = \text{máx}\{u, 1-u\}$. Calcule la distribución acumulada de la variable $X = h(U)$. Calcule $E(X)$ y $V(X)$.

Solución:

$$P(\text{máx}\{U, 1-U\} \leq x) = P(U \leq x, 1-U \leq x) = P(1-x \leq U \leq x) = 2x\mathbf{1}\{1/2 \leq x \leq 1\}. \tag{12}$$

Independencia Dado un vector (X,Y) (continuo o discreto) decimos que las variables X e Y son independientes si

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B), \quad \text{para todo } A, B \subset \mathbb{R}.$$

Lema 9. *Si X e Y son discretas y vale*

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \textit{para } (x,y) \textit{ en el rango de } (X,Y),$$

entonces X e Y son independientes.

Si X e Y son continuas y vale

$$f_{(X,Y)}(x,y) = f_X(x)f_Y(y) \quad \textit{para todo } x, y \textit{ en } \mathbb{R},$$

entonces X e Y son independientes.

Ejemplo Tiramos una moneda 2 veces $X = 1$ si el número de caras es par. $Y = 1$ si la primera moneda es cara. Pregunta: X e Y son independientes? Calculamos primero las marginales:

$$P(X = 0) = P(X = 1) = 1/2, \qquad P(Y = 0) = P(Y = 1) = 1/2 \tag{13}$$

Ahora la conjunta:

$$\begin{aligned} P(X = 0, Y = 1) &= P[\text{primera cara y número par de caras}] \\ &= P(\{(1,1)\}) = 1/4 \end{aligned} \tag{14}$$

Esto es suficiente para probar que X e Y son independientes, porque como las variables asumen solamente dos valores, se puede usar que A, B independientes implica A, B^c independientes, para obtener que las otras probabilidades conjuntas también son iguales a los productos de las marginales correspondientes.

Ejemplo Sea $S :=$ espacio de las permutaciones de $\{1,2,3,4\}$ y sea $X = (X_1, \ldots, X_4)$ una permutación aleatoria: $P(X = \pi) = \frac{1}{4!}$. Sea $Y = \mathbf{1}\{X_3 = 4\}$ y $Z = \mathbf{1}\{X_1 > X_2\}$. Demuestre que Y y Z son independientes. Hay 3! permutaciones posibles con $X_3 = 4$; la mitad de esas permutaciones tiene $X_1 < X_2$ y la otra mitad la desigualdad contraria. Por lo tanto

$$\begin{aligned} P(Z=1, Y=1) = P(X_1 < X_2, X_3 = 4) &= \frac{3!/2}{4!} \\ &= \frac{1}{2}\frac{1}{4} = P(X_1 < X_2)P(X_3 = 4) = P(Z=1)P(Y=1). \end{aligned} \tag{15}$$

Concluimos que el orden relativo de las coordenadas 1 y 2 es independiente del valor de la coordenada 4.

Lema 10. *Si existen h y g tales que*

$$P(X = x, Y = y) = Ch(x)g(y), \quad \textit{para todo } x, y$$

entonces X e Y son independientes y sus distribuciones marginales son

$$P(X = x) = h(x)\big(\sum_{x'} h(x')\big)^{-1}, \qquad P(Y = y) = g(y)\Big(\sum_{y'} g(y')\Big)^{-1}.$$

Dem: Como la suma sobre todos los valores tiene que ser 1, la constante C tiene que ser

$$C = \Big(\sum_x h(x) \sum_y g(y)\Big)^{-1}$$

Sumando sobre y tenemos la marginal de X:

$$P(X = x) = Ch(x) \sum_{y'} g(y') = h(x)\big(\sum_{x'} h(x')\big)^{-1}$$

y sumando sobre x:

$$P(Y = y) = Cg(y) \sum_{x'} h(x') = g(y)\Big(\sum_{y'} g(y')\Big)^{-1},$$

Así:

$$\begin{aligned} P(X = x)P(Y = y) &= h(x)g(y)\Big(\sum_{y'} g(y') \sum_{x'} h(x')\Big)^{-1} \\ &= Ch(x)g(y) = P(X = x, Y = y). \quad \square \end{aligned}$$

Ejemplo Sea (X, Y) un vector aleatorio con distribución conjunta

$$p(k, \ell) = C\frac{\lambda^k \mu^\ell}{k!\ell!}$$

$k, \ell = 0, 1, 2, \ldots;\ \lambda, \mu > 0$.

Claramente $p(k, \ell) = C\, g(k) f(\ell)$, por lo tanto son independientes. Calculemos C:

$$C = \Big(\sum_k \frac{\lambda^k}{k!} \sum_\ell \frac{\mu^\ell}{\ell!}\Big)^{-1} = e^{-\lambda} e^{-\mu}. \tag{16}$$

La marginal de X es

$$P(X = k) = \sum_{\ell \geq 0} \frac{\lambda^k \mu^\ell\, e^{-\lambda-\mu}}{k!\ell!} = \frac{\lambda^k\, e^{-\lambda}}{k!} \sum_{\ell \geq 0} \frac{\mu^\ell\, e^{-\mu}}{\ell!} = \frac{\lambda^k\, e^{-\lambda}}{k!}$$

Es decir, $X \sim$ Poisson(λ). Similarmente $Y \sim$ Poisson(μ).

Ejemplo (X, Y) tiene distribución conjunta

$$p(k, n) = C\frac{2^{-k}}{n}, \quad k = 1, 2, \ldots;\ n = 1, \ldots, k$$

C constante apropiada.

Como $p(k,n) = C2^{-k}\frac{1}{n}$, parecería que $p(k,n)$ puede factorizarse; esto implicaría que X, Y serían independientes.

Pero no. Hay dependencia entre X e Y porque

$$p(k,n) = C\frac{2^{-k}}{n}\mathbf{1}\{n \leq k\}$$

no se puede factorizar. Así que X e Y **no** son independientes.

Esta conclusión se puede obtener de los siguientes cálculos:

$$P(X=1) > 0,\ P(Y=2) > 0, \quad P(X=1, Y=2) = 0.$$

Independencia de variables aleatorias continuas

X e Y son *independientes* si y solo si para todo x, y,

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

Lema *las variables continuas X e Y con densidad f_X, f_Y, respectivamente son independientes si y sólo si*

$$f_X(x)f_Y(y) = f(x,y),\ \textit{para todo } x, y$$

Dem: Ejercicio.

Ejemplo X Y con densidad conjunta $f(x,y) = e^{-x-y}$, $x, y > 0$. Entonces $f(x,y)$ se factoriza como $f(x,y) = e^{-x}e^{-y}$ y son independientes.

Def Una familia $(X_i : i \in J)$ de *vectores* aleatorios es independiente (mutuamente independientes) si para todo subconjunto finito de índices $K \subset J$,

$$P(X_i \leq a_i,\ i \in K) = \prod_{i \in K} P(X_i \leq a_i), \quad \forall a_i \in \mathbb{R}$$

Ejemplos

1. Encuentros casuales. Dos personas deciden encontrarse un día entre las 5 y las 6. Cada uno llega en instantes independientes distribuidos uniformemente en ese intervalo y espera 15 minutos. Cual es la probabilidad que se encuentren?

El vector (X, Y) representa los tiempos de llegadas de las dos personas. Ese vector está distribuido uniformemente en $[0, 60]^2$.

Definimos la región en la cual se encuentran: la distancia entre llegadas tiene que ser menor que 15:

$$A := \{(x,y) \in [0,60]^2 : |x - y| \leq 15\}$$

queremos calcular $P((X,Y) \in A)$, con (X,Y) uniforme en $[0,60]^2$:

$$f(x,y) = \frac{1}{60^2}\mathbf{1}\{(x,y) \in [0,60]^2\}$$

$$P((X,Y) \in A) = \frac{\text{area}(A)}{60^2} = 1 - \frac{\text{area}(A^c)}{60^2} = 1 - \frac{45^2}{60^2} = \frac{7}{9}$$

2. Permutaciones. Sean $X_1, \ldots, X_n$ una familia de n variables continuas independientes con densidad común f y acumulada F estrictamente creciente. Demuestre que la familia $(F(X_1), \ldots, F(X_n))$ es una familia de variables uniformes en $[0,1]$ independientes.

Respuesta: Las marginales son uniformes: para $x \in (0,1)$,

$$P(F(X_i) \leq x) = P(X_i \leq F^{-1}(x)) = F(F^{-1}(x)) = x,$$

porque F es inversible en ese intervalo. Esto implica que $F(X_i) \sim \text{Uniforme}[0,1]$.

Sean $S_1, \ldots, S_n$ las estadísticas de orden definidas por

$$S_1 < \ldots < S_n; \quad \{X_1, \ldots, X_n\} = \{S_1, \ldots, S_n\} \text{ (como conjuntos)}$$

es decir, $S_1 = \text{mín}_i\, S_i$, $S_n = \text{máx}_i\, S_i$, etc.

Sea $K(i)$ el lugar de X_i cuando las variables son ordenadas: $X_i = S_{K(i)}$.

Muestre que $(K(1), \ldots, K(n))$ es una permutación aleatoria de $(1, \ldots, n)$.

Respuesta: Como las variables son independientes e identicamente distribuidas, vale que

$$P(X_1 < X_2 < \ldots < X_n) = P(X_{\pi(1)} < \ldots < X_{\pi(n)})$$

para cualquier permutación π de $(1, \ldots, n)$. Por lo tanto, $P(K = \pi) = 1/n!$

3. Secretarias. Hay n candidatas a secretaria, con puntajes $Y_1, \ldots, Y_n$, variables continuas identicamente distribuidas e independientes en $\mathbb{R}^+$. Como desconocemos la distribución de Y_n sólo podemos saber el orden relativo de las candidatas entrevistadas. Después de entrevistar una candidata, podemos decidir contratarla o no. Candidata entrevistada y no contratada inmediatamente deja de ser candidata. Queremos encontrar una política que maximice la probabilidad de elegir la mejor candidata.

Política r: entrevistar $r-1$ candidatas sin contratar ninguna, seguir entrevistando y contratar la primera que sea mejor que todas las anteriores (o la última, en caso que la mejor de todas esté entre las primeras $r-1$ entrevistadas).

Sea $p(r) :=$ probabilidad de encontrar la mejor candidata con la política r.

Buscamos el r que maximiza $p(r)$. Por probabilidad total tenemos

$$\begin{aligned} p(r) &= \sum_{i=1}^{n} P(\text{candidata } i \text{ es seleccionada, } i \text{ es la mejor candidata}) \\ &= \sum_{i=r}^{n} P(i \text{ es seleccionada, } i \text{ es la mejor}) \end{aligned}$$

porque las $r-1$ primeras candidatas tienen probabilidad cero de ser seleccionadas. Definamos el evento

$$\begin{aligned} A_i := \ & \text{la mejor entre las } i-1 \text{ primeras entrevistadas} \\ & \text{está entre las } r-1 \text{ primeras entrevistadas} \end{aligned}$$

Vale la siguiente identidad de eventos para $i \geq r$:

$$\{i \text{ seleccionada, } i \text{ es la mejor}\} = A_i \cap \{i \text{ es la mejor}\}$$

de donde

$$p(r) = \sum_{i=r}^{n} P(A_i \cap \{i \text{ es la mejor}\}) \tag{17}$$

El evento A_i depende del orden relativo entre las primeras $i-1$ candidatas que es independiente del evento $\{i \text{ es la mejor}\}$; esto se demuestra con un argumento parecido al de (15). Como $P(A_i) = (r-1)/(i-1)$ y $P(i \text{ es la mejor}) = 1/n$, tenemos

$$p(r) = \sum_{i=r}^{n} \frac{r-1}{i-1}\frac{1}{n} = \frac{r-1}{n}\sum_{i=r}^{n}\frac{1}{i-1}$$

La suma no está definida para $r = 1$ pero en ese caso $p(r) = \frac{1}{n}$.

Si pensamos que n es grande y que la solución será una fracción de n, es decir $r = xn$, para algún $x \in (0,1)$,

$$p(xn) \approx \frac{xn}{n}\sum_{i=xn}^{n}\frac{1}{i} = x\frac{1}{n}\sum_{i=xn}^{n}\frac{1}{i/n} \approx x\int_x^1 \frac{1}{t}dt = -x\log x,$$

aproximando la suma por la integral. Derivando e igualando a 0:

$$-\log x + 1 = 0. \tag{18}$$

cuya solución es $x = e^{-1}$. Por lo tanto el máximo se obtiene cuando $x = e^{-1}$ y la probabilidad de seleccionar la mejor candidata usando la política óptima $r = e^{-1}n$ es aproximadamente

$$p(e^{-1}n) \approx -e^{-1}\log e^{-1} = e^{-1} = 0{,}36787944117\ldots$$

4. Records. Sean $X_1, X_2, \ldots$ una familia de variables continuas independientes. Sea $Y_n = \mathbf{1}\{X_n > X_i,$ para todo $1 \leq i < n\}$. Es decir que $Y_n = 1$ si hay un record en el instante n. Pregunta: $Y_1, Y_2, \ldots$ son variables independientes?

Basta probar que las primeras n son independientes.

5. Aguja de Buffon En un piso de tabla corrida, las lineas determinadas por las tablas son paralelas y están a distancia D. Una aguja de longitud $L < D$ es lanzada al azar sobre ese piso y se considera el evento $A =$ "la aguja interseca una de las lineas". El evento complementario es $A^c =$ "la aguja está totalmente dentro de una de las tablas".

Veremos que la probabilidad de A depende del número π. Las variables relevantes son:

$X =$ distancia del centro de la aguja a la paralela más cercana

$\theta =$ ángulo entre la recta que contiene la aguja y la recta perpendicular a las tablas que contiene el centro de la aguja.

$X \sim$ Uniforme$[0, D/2]$. $f_X(x) = \frac{2}{D}\mathbf{1}\{x \in [0, d/2]\}$.

$\theta \sim$ Uniforme$[0, \pi/2]$. $f_\theta(y) = \frac{2}{\pi}\mathbf{1}\{y \in [0, \pi/2]\}$.

X y θ son independientes.

La aguja interseca una de las paralelas si

$$X < \frac{L}{2}\cos\theta,$$

que equivale a

$$(X, \theta) \in \Big\{(x, y) \in \Big[0, \frac{D}{2}\Big] \times \Big[0, \frac{\pi}{2}\Big] : x < \frac{L}{2}\cos y\Big\}$$

$$= \Big\{(x, y) : 0 < y < \frac{\pi}{2},\ 0 < x < \frac{L}{2}\cos y\Big\}$$

Entonces

$$P(A) = P\Big(X < \frac{L}{2}\cos\theta\Big) = \int_0^{\pi/2}\int_0^{\frac{L}{2}\cos y} f_X(x) f_\theta(y)\,dx\,dy$$

$$= \frac{4}{\pi D}\int_0^{\pi/2}\int_0^{\frac{L}{2}\cos y} dx\,dy = \frac{4}{\pi D}\int_0^{\pi/2}\frac{L}{2}\cos y\,dy = \frac{2L}{\pi D}$$

Esto se usa para "estimar" π usando

$$\pi = \frac{2L}{P(A)D}$$

Llamemos $p = P(A)$. Repitiendo el experimento muchas veces y tomando la proporción muestral $\hat{p}$ de éxitos, se estima π por $\hat{\pi} = \frac{2L}{\hat{p}D}$.

Esperanza de funciones de vectores

$$Eh(X, Y) = \int\int h(x, y) f(x, y)\,dx\,dy$$

$$E(aX + bY) = aEX + bEY$$

Si X e Y son independientes:

$$E(XY) = EX\,EY$$

Contraejemplo de funciones con $EXY = EX\,EY$ pero no son independientes:

$$f(x,y) = C\mathbf{1}\{x^2 + y^2 \leq 1\}$$

No son independientes porque el rango del vector no es el producto cartesiano de los rangos. La esperanza de cada variable es 0 y también lo es la esperanza del producto.

El juego de los sobres.

Dos sobres con plata Y_1, Y_2. iid Uniformes en $[0, 10]$.

Abro un sobre y veo y. Debo cambiar de sobre? Quiero maximizar la esperanza del valor que voy a ganar.

Estrategia 1: Fijo $K \in (0, 10)$. Si $y > K$, me quedo con y. Si no, cambio.

Sea X_1 valor del primer sobre.

X_2 valor obtenido despues de aplicar la estrategia.

$X_2 = Y_1\mathbf{1}\{Y_1 > K\} + Y_2\mathbf{1}\{Y_1 \leq K\}$

$EX_2 = E(Y_1\mathbf{1}\{Y_1 > K\}) + EY_2P(Y_1 \leq K)$

$= \int_K^{10} yf(y)dy + 5\,P(Y \leq K) = \left[\frac{x^2}{2\,10}\right]_K^{10} + 5\frac{K}{10}$

$= 5 - \frac{K^2}{2\,10} + 5\frac{K}{10} = 5 + \frac{K}{10}(5 - \frac{K}{2})$

EX_2 asume un máximo en $K = 5$.

Para verlo, multiplique por 2 y vea que $g(K) = K(10 - K)$ es una parabola con inclinación para abajo que pasa por 0 y 10, por lo tanto asume su máximo en 5.

En resumen, la estrategia queda:

Miro Y_1, si es mayor que 5, me quedo. Si no, me paso a Y_2.

La media para $K = 5$ queda $EX_2 = 6{,}25$

Covarianza y correlación Sean X e Y dos v.a. con esperanzas EX y EY respectivamente, la *covarianza* entre X e Y se define como

$$E(X - EX)(Y - EY) = \text{caso continuo y discreto}$$

Observación: C$ov(X, X) = V(X)$.

Idea intuitiva: Si X e Y tienen una fuerte relación positiva, en el sentido que valores grandes de X aparecen asociados con valores grandes de Y y valores pequeños de X aparecen asociados con valores pequeños de Y, entonces los productos serán positivos y por lo tanto la covarianza será positiva.

Por otra parte, si X e Y tienen una fuerte relación negativa, en el sentido que valores grandes de X aparecen asociados con valores pequeños de Y y valores pequeños de X aparecen asociados con valores grandes de Y, entonces la mayoría de los productos serán negativos y por lo tanto la covarianza será negativa.

Propo C$ov(X, Y) = E(XY) - EX\,EY$.

Probarlo para discreto. Continuo igual.

Ejemplo discreto:

	0	1	2	X
0	0.4	0.1	0.1	0.6
1	0.1	0.2	0.1	0.4
Y	0.5	0.3	0.2	1

Ejemplo continuo: $f(x, y) = \frac{6}{5}(x + y^2)\mathbf{1}\{(x, y) \in [0, 1]^2\}$.

C$ov(X, Y) = -\frac{1}{100}$

Propo Si X e Y son independientes, C$ov(X, Y) = 0$. La reciproca no es verdadera.

Dem Como las variables son independientes las funciones de probabilidad en el caso discreto y las densidades en el caso continuo factorizan. Por ejemplo en el caso continuo.

$$EXY = \int_{\mathbb{R}^2} xyf_X(x)f_Y(y)dxdy = \int_{\mathbb{R}} xf_X(x)dx \int_{\mathbb{R}} yf_Y(y)dy$$

Contraejemplo: X e Y tienen covarianza cero pero no son indep:

	-1	0	1	X
-1	1/8	0	1/8	1/4
0	0	1/2	0	1/2
1	1/8	0	1/8	1/4
Y	1/4	1/2	1/4	1

Ejercicio: Contraejemplo continuo Buscar una densidad que satisfaga: $f(x,y) = f(x,-y) = f(-x,y) = f(-x,-y)$ que garantiza que $E(XY) = 0$ y $EX = EY = 0$ pero que no sea el producto de dos funciones.

Verifique que por ejemplo $f(x,y)$ uniforme en una bola centrada en 0 satisface.

Coeficiente de correlación Sean X e Y dos v.a. con esperanzas EX y EY respectivamente y varianza positiva, el coeficiente de correlación entre X e Y se define como

$$\rho(X,Y) = \frac{\mathrm{C}ov(X,Y)}{\sigma_X \sigma_Y}$$

Proposición 11. *1. Sean a, b, c y d números reales, $a \neq 0$, $c \neq 0$ y X e Y v.a. con varianza positiva, entonces*

$$\rho(aX+b, cY+d) = \mathrm{s}g(ac)\rho(X,Y)$$

donde sg denota la función signo.

2. $-1 \leq \rho(x,y) \leq 1$

3. $|\rho(X,Y)| = 1$ *sii* Y *es funcion lineal de* X*.*

Dem: 1. Cuentas.

2. Asumamos $EX = EY = 0$.

Defina $g(t) = E(X - tY)^2$

Claramente $g(t) \geq 0$

$$g(t) = EX^2 - 2t\,E(XY) + t^2 EY^2$$

Polinomio de segundo grado en t. $a = EY^2$, $b = -2E(XY)$, $c = EX^2$.

Discriminante $b^2 - 4ac = 4(E(XY))^2 - 4EX^2EY^2 \leq 0$

Por lo tanto

$$\frac{(E(XY))^2}{EX^2EY^2} \leq 1$$

es decir $\rho^2 \leq 1$, lo que implica $-1 \leq \rho \leq 1$.

Caso general: basta ver que $\rho(X,Y) = \rho(X - EX, Y - EY)$.

3. Supongamos que $\rho = 1$. Esto implica que el discriminante de $g(t)$ es cero y que g tiene una única raiz t_0. Es decir

$$E(X - t_0 Y)^2 = 0$$

Como X e Y tienen esperanza cero, $X - t_0 Y = 0$ con probabilidad 1.

Caso general, substituyendo

$$E(X - EX - t_0(Y - EY))^2 = 0$$

implica que $Y = \frac{1}{t_0}X + \frac{1}{t_0}EY - EX$.

Reciprocamente, si $Y = AX + B$ entonces $|\rho| = 1$ (cuenta).

Distribución condicional Dado vector (X, Y), La distribución condicional de X dado Y está dada por

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Esperanza condicional

$$E(X|Y = y) = \sum_x xP(X = x|Y = y)$$

Teorema 12. *Vale*

$$E(X) = \sum_y E(X|Y = y)P(Y = y)$$

Teorema de la esperanza total $A_1, \ldots, A_n$ partición del espacio muestral. B evento.

Teorema de la proba total: $P(B) = \sum_i P(B|A_i)P(A_i)$

$$P(X = k) = \sum_i P(X = k|A_i)P(A_i)$$

Lema $EX = \sum_i E(X|A_i)P(A_i)$

Dem $EX = \sum_k kP(X = k) = \sum_k k \sum_i P(X = k|A_i)P(A_i)$

$$\sum_i \sum_k kP(X = k|A_i)P(A_i) = \sum_i E(X|A_i)P(A_i) \qquad \square$$

Ejemplo Cálculo de la esperanza de la geométrica usando esperanza total. Si condicionamos al resultado del primer ensayo:

$$EX = E(X|X = 1)P(X = 1) + E(X|X > 1)P(X > 1)$$

Claramente, $E(X|X = 1) = 1$ y por lo que calculamos arriba, $E(X|X > 1) = EX + 1$. Como $P(X = 1) = p$,

$$EX = 1p + (EX + 1)(1 - p) = p + EX - pEX + 1 - p \Longrightarrow EX = 1/p$$

Ejemplo Gallina produce N huevos Poisson λ. Cada huevo produce un pollo con proba p independiente de los otros. Sea K el número de pollos.

Calcule $E(K|N = n)$ y $E(K)$.

Note que

$$P(K = k|N = n) = \binom{n}{k} p^n (1 - p)^{n-k}$$

Asi

$$E(K|N = n) = np$$

$$EK = \sum_n E(K|N = n)P(N = n) = \sum_n npP(N = n) = pEN = \lambda p$$

Se puede calcular tambien $P(K = k)$ directamente.

Se puede calcular $P(N = n|K = k)$ y $E(N|K = k)$.

Juego de los sobres Dos sobres. Uno contiene a pesos y el otro b pesos; $a < b$. Desconocemos los valores a y b.

Usted elije uno de los sobres, lo abre y observa el valor que contiene.

Le ofrezco la oportunidad de elegir el otro sobre.

Tiene sentido cambiarse de sobre?

Más precisamente: hay un estrategia que le permita elegir el sobre con b pesos con proba estrictamente mayor que 1/2?

Estrategia: Sea X_1: valor en el sobre elegido.

$P(X_1 = a) = P(X_1 = b) = 1/2$

Sea $Y \sim$ exponencial(1), una variable independiente de X_1

Observe X_1 y simule Y.

Si $X_1 < Y$ cambie de sobre; si $X_1 > Y$ no cambie.

X_2: valor en el sobre final (después de un eventual cambio).

Sabemos calcular las probabilidades condicionales siguientes:

$P(X_2 = b|X_1 = b) = P(Y < b) = 1 - e^{-b}$,

$P(X_2 = b|X_1 = a) = P(Y > a) = e^{-a}$.

Usando el teorema de la probabilidad total:

$$\begin{aligned} &P(X_2 = b) \\ &\quad = P(X_2 = b|X_1 = b)P(X_1 = b) + P(X_2 = b|X_1 = a)P(X_1 = a) \\ &\quad = \frac{1}{2}(1 - e^{-b}) + \frac{1}{2}e^{-a} = \frac{1}{2} + \frac{1}{2}(e^{-a} - e^{-b}) > \frac{1}{2} \end{aligned}$$

Problema del minero Un minero está en el fondo de una mina y ve tres túneles: 1, 2 y 3. El tunel 1 lleva a la salida en una hora. El tunel 2 vuelve a la misma encrucijada en 2 horas y el tunel 3 vuelve a la encrucijada en 3 horas. Cada vez que el minero está en la encrucijada, elige uno de los túneles con probabilidad 1/3, independientemente de lo que eligió antes.

Sea T el tiempo que tarda en salir de la mina. T es finito con probabilidad 1:

$$P(T = \infty) \leq \Big(\frac{2}{3}\Big)^n \quad \text{para todo } n \geq 0.$$

Por otro lado $T \leq T_1$ el tiempo que se tarda si todos los túneles demoran 3 horas. Pero $T_1 = 3G$, donde G es una geométrica de parámetro 1/3. Así $ET \leq ET_1 = 3/(1/3) = 9$. Esto es una cota.

Calculemos exactamente ET. Sea X el tunel que el minero elige en la primera tentativa. Entonces

$$E(T|X = 1) = 1, \quad E(T|X = 2) = 2 + ET, \quad E(T|X = 3) = 3 + ET$$

Usando esto, calculamos

$$\begin{aligned} ET = E(E(T|X)) &= \frac{1}{3}E(T|X = 1) + \frac{1}{3}E(T|X = 2) + \frac{1}{3}E(T|X = 3) \\ &= 1\frac{1}{3} + (2 + ET)\frac{1}{3} + (3 + ET)\frac{1}{3} = 2 + \frac{2}{3}ET. \end{aligned}$$

Despejando ET, obtenemos $ET = 6$. Usamos que $ET < \infty$ para poder pasar de término.

Se puede resolver tambien particionando con $K :=$ número de intentos hasta encontrar el camino 1. K es una variable aleatoria Geométrica(1/3): $P(K = k) = (2/3)^{k-1}(1/3)$, $k \geq 1$.

Como $T = \sum_k T\mathbf{1}\{K = k\}$,

$$E(T\mathbf{1}\{K = k\}) = E\Big(\sum_{i=1}^{k} X_i\mathbf{1}\{K = k\}\Big) = \sum_{i=1}^{k} E(X_i\mathbf{1}\{K = k\}).$$

Como $\{K = k\} = \{X_1 \in \{2, 3\}, \ldots, X_{k-1} \in \{2, 3\}, X_k = 1\}$, tenemos que

$$E(X_k\mathbf{1}\{K = k\}) = 1P(K = k) \tag{19}$$

y para $1 \leq i < k$,

$$\begin{aligned} E(X_i\mathbf{1}\{K = k\}) &= 2P(X_i = 2, K = k) + 3P(X_i = 3, K = k) \\ &= 2\frac{1/3}{2/3}P(K = k) + 3\frac{1/3}{2/3}P(K = k) = \frac{5}{2}P(K = k) \end{aligned} \tag{20}$$

Por lo tanto,

$$ET = \sum_k E(T\mathbf{1}\{K=k\}) = \sum_k (1+(k-1)\frac{5}{2})P(K=k) = (1+2\frac{5}{2}) = 6.$$

Distribución condicional de variables continuas

(X,Y) vector aleatorio con densidad f.

Queremos definir $P(Y \leq y | X = x)$

Si X es continua, $P(X=x)=0$. Procedimiento límite:

$$\star := P(Y \leq y | x \leq X \leq x+h) = \frac{P(Y \leq y, x \leq X \leq x+h)}{P(x \leq X \leq x+h)}$$

$$= \frac{\int_{-\infty}^{y}\int_{x}^{x+h} f(u,v)dudv}{\int_x^{x+h} f_X(v)dv}$$

dividiendo arriba y abajo por h y sacando límite,

$$\lim_{h\to 0} \star = \int_{-\infty}^{y} \frac{f(x,v)}{f_X(x)}dv$$

Así definimos $f_{Y|X=x}(y) = f(x,y)/f_X(x)$ para x tal que $f(x) \neq 0$.

$f_{Y|X=x}$ es una densidad: $\int f_{Y|X=x}(y)dy = \int \frac{f(x,y)}{f_X(x)}dy = 1$.

Es la densidad de una nueva variable con esperanza:

$$E(Y|X=x) = \int_{-\infty}^{\infty} y\, f_{Y|X=x}(y)dy$$

Valen las siguientes fórmulas:

$$P(Y \leq y) = \int_{-\infty}^{\infty} P(Y \leq y | X = x) f_X(x)dx$$

$$EY = \int_{-\infty}^{\infty} E(Y|X=x) f_X(x)dx$$

Ejemplos

1. Sea (X,Y) un vector continuo en $\mathbb{R}^2$ con densidad

$$f(x,y) = \lambda^2 e^{-\lambda y}\mathbf{1}\{0 \leq x \leq y\} \tag{21}$$

Calculemos primero la marginal de Y:

$$f_Y(y) = \int f(x,y)dx = \int_0^{\infty} \lambda^2 e^{-\lambda y}\mathbf{1}\{0 \leq x \leq y\}dx = \lambda^2 y e^{-\lambda y}\mathbf{1}\{y>0\},$$

es decir que $Y \sim \text{Gama}(2,\lambda)$. La condicionada de X dado $Y=y$ es

$$f_{X|Y=y}(x) = \frac{\lambda^2 e^{-\lambda y}}{\lambda^2 y e^{-\lambda y}}\mathbf{1}\{0<x<y\} = \frac{1}{y}\mathbf{1}\{0<x<y\}$$

es decir, Uniforme$[0,y]$. La marginal de X es

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{\infty} \lambda^2 e^{-\lambda y}\mathbf{1}\{0 \leq x \leq y\}dy$$
$$= \int_x^{\infty} \lambda^2 e^{-\lambda y}dy = \lambda e^{-\lambda x}\mathbf{1}\{x \geq 0\}$$

O sea, $X \sim$ Exponencial(λ). Finalmente, sea $Z = Y - X$. La distribución conjunta de (X, Z) está dada por

$$\begin{aligned} P(X > x, Z > z) &= P((X,Y) \in \{(x', y') : x' > x,\, y' > z + x'\}) \\ &= \int_x^\infty \int_{x'+z}^\infty \lambda^2 e^{-\lambda y'} dy' dx' \\ &= \lambda \int_x^\infty e^{-\lambda(x'+z)} dx' = e^{-\lambda x} e^{-\lambda z}. \end{aligned}$$

o sea, X y Z son Exponencial(λ) independientes.

Con esto en manos, se puede ver que la distribución de Y dado $X = x$ es la misma que la de $Z + x$ con $Z \sim$ Exponencial(λ). Por lo tanto las esperanzas condicionadas se calculan así.

$$E(Y|X = x) = \frac{1}{\lambda} + x \quad \Rightarrow \quad E(Y|X) = \frac{1}{\lambda} + X$$

Por otro lado,

$$E(X|Y = y) = \frac{y}{2} \quad \Rightarrow \quad E(X|Y) = \frac{Y}{2}$$

Controlando las esperanzas de las marginales da todo bien:

$$EY = E(E(Y|X)) = E\Big(\frac{1}{\lambda} + X\Big) = \frac{1}{\lambda} + EX = \frac{2}{\lambda}$$

$$EX = E(E(X|Y)) = E(Y/2) = \frac{1}{2}EY = \frac{1}{2}\frac{2}{\lambda} = \frac{1}{\lambda}.$$

porque Y era una Gamma$(2, \lambda)$ y X una Exponencial(λ).

Note finalmente que si $Y \sim$ Gama$(2, \lambda)$ y $U \sim$ Uniforme$[0, 1]$, entonces el vector (UY, Y) tiene distribución (21). Para verlo, observe que dado $Y = y$ la distribución de UY es uniforme en $[0, y]$. Por lo tanto la densidad conjunta de (UY, Y) es

$$\frac{1}{y}\lambda^2 y e^{-\lambda y}\, \mathbf{1}\{0 \le x \le y\}$$

y la densidad conjunta de $(X, Z) = (UY, (1 - U)Y)$ es

$$f_{X,Z}(x, z) = \lambda^2 e^{-\lambda x} e^{-\lambda z}\, \mathbf{1}\{x \ge 0, z \ge 0\}.$$

2. $f(x, y) = 2(x + 2y)I_T(x, y)$ con $T = \{0 \le x \le 1, 0 \le y \le 1 - x\}$

Calcular las marginales de X e Y.

$f_X(x) = 2(1 - x)I_{[0,1]}(x)$

$f_Y(y) = (1 + 2y - 3y^2)I_{[0,1]}(y)$

Calcular $P(X \le 1/2|Y \le 1/4) = 8/19$

$P(X \le 1/2|Y = 1/4) = \int_0^{1/2} \frac{f(x,1/4)}{f_Y(1/4)} dx$

Densidad condicional e Independencia

X e Y son indep si $f(x, y) = f_X(x)f_Y(y)$.

En función de proba condicional:

$$f_X(x) = f_{X|Y=y}(x)$$

Dem: Por la def de la densidad condicional, $f(x, y) = f_Y(y)f_{X|Y=y}(x)$.

Por lo tanto las variables son independientes si y solo si $f_X(x) = f_{X|Y=y}(x)$

Para probar que dos variables continuas **no son independientes** basta exhibir un rectangulo $[a, b]x[c, d]$ tal que

$$\int_a^b \int_c^d f(x, y)dxdy \neq \int_a^b f_X(x)dx \int_c^d f_Y(y)dy$$

Si $R_{X,Y} \neq R_X \times R_Y$, las variables no son independientes.

Otra forma de probar que X e Y no son independientes es encontrar un punto (u, v) en $\mathbb{R}^2$ tal que $f(x, y)$, $f_X(x)$ y $f_Y(y)$ sean todas continuas en ese punto y$f(x, y) \neq f_X(x)f_Y(y)$.

Por continuidad, la condición se cumplirá en un entorno rectangular del punto.

Distribución de la suma de dos variables

Caso discreto

Sea (X, Y) un vector aleatorio discreto con distribución conjunta p y sea $Z = X + Y$. La distribución de Z es

$$P_Z(z) = \sum_x p_{X,Y}(x, z - x) = \sum_y p_{X,Y}(z - y, y))$$

Cuando X e Y son independientes,

$$P_Z(z) = \sum_x p_Y(z - x)p_X(x) = \sum_y p_X(z - y)p_Y(y)$$

Aplicación: suma de Poisson independientes es Poisson: Sea $X \sim$ Poisson(λ), $Y \sim$Poisson(μ). Entonces, $X + Z \sim$ Poisson($\lambda + \mu$).

$$P(Z = n) = \sum_{k=0}^{n} p_X(k)p_Y(n - k) = \sum_{k=0}^{n} \frac{e^{-\lambda}\lambda^k}{k!} \frac{e^{-\mu}\mu^{n-k}}{(n-k)!}$$

$$= \frac{e^{-(\lambda+\mu)}(\lambda + \mu)^n}{n!} \sum_{k=0}^{n} \binom{n}{k} \Big(\frac{\lambda}{\lambda + \mu}\Big)^k \Big(\frac{\mu}{\lambda + \mu}\Big)^{n-k}$$

Caso continuo X Y va continuas con f. $Z = X + Y$. Entonces

$$P(Z \leq z) = \int \int_{\{(x,y):x+y\leq z\}} f(x, y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y)dxdy$$

substituya $u = x$, $v = y + x$:

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(u, v - u)dudv$$

de donde

$$f_Z(z) = \int_{-\infty}^{\infty} f(x, z - x)dx$$

Caso independiente:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx$$

La densidad de la suma de dos variables independientes es la convolución de las densidades de las variables.

Gama $X_1, \ldots, X_n$ exponenciales indep. $Z_n = X_1 + \cdots + X_n$. Entonces

$$f_Z(z) = \frac{\lambda^n}{(n-1)!} z^{n-1} e^{-\lambda z} \qquad Gama(n, \lambda)$$

Inducción. Suponga que $T = X_1 + \cdots + X_{n-1}$ es Gama$(n - 1, \lambda)$. Como T y X_n son independientes:

$$\begin{aligned} f_Z(z) &= \int_0^z \frac{\lambda^{n-1}}{(n-2)!} x^{n-2} e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx \\ &= \frac{\lambda^n}{(n-2)!} e^{-\lambda z} \int_0^z x^{n-2} dx \\ &= \frac{\lambda^n}{(n-1)!} e^{-\lambda z} x^{n-1} \end{aligned} \tag{22}$$

9. Ley de grandes números

Esperanzas de funciones de variables aleatorias

Se aplican las fórmulas siguientes que se pueden probar como lo hicimos para el caso de una variable:

Caso discreto:

$$Eg(X_1,\dots,X_n) = \sum_{x_1,\dots,x_n} g(x_1,\dots,x_n)P(X_1 = x_1,\dots,X_n = x_n)$$

(si la suma está bien definida)

Caso continuo. Vector $(X_1,\dots,X_n)$ con densidad conjunta f.

$$Eg(X_1,\dots,X_n) = \int_{\mathbb{R}^n} g(x_1,\dots,x_n)f(x_1,\dots,x_n)\,dx_1,\dots,dx_n$$

(si la integral está bien definida)

Las fórmulas valen también para vectores infinitos (si las sumas e integrales están bien definidas).

Esperanzas y varianzas de sumas de variables aleatorias

$$E\big(\sum_i a_iX_i\big) = \sum_i a_i\,EX_i$$

$$V\big(\sum_i a_iX_i\big) = \sum_i a_i^2VX_i + 2\sum_{i<j} a_ia_j\mathrm{Cov}(X_i,X_j)$$

Si son independientes, como las covarianzas son 0,

$$V\big(\sum_i a_iX_i\big) = \sum_i a_i^2VX_i$$

Muestra. Una *muestra* de una variable aleatoria X es un vector $X_1,\dots,X_n$ de variables aleatorias independientes identicamente distribuidas (iid) con $X_i \sim X$.

Defina la **media muestral** de una muestra por

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$$

Si $EX = \mu$ y $VX = \sigma^2$, obtenemos

$$E\overline{X}_n = \mu, \qquad V\overline{X}_n = \sigma^2/n$$

Desigualdad de Markov ***Sea*** $X \geq 0$ ***una variable aleatoria no negativa con esperanza finita y*** $\varepsilon > 0$***. Entonces***

$$P(X \geq \varepsilon) \leq \frac{EX}{\varepsilon}$$

Dem

$$X = X\,\mathbf{1}\{X < \varepsilon\} + X\,\mathbf{1}\{X \geq \varepsilon\} \;\geq\; \varepsilon\mathbf{1}\{X > \varepsilon\}$$

porque como $X \geq 0$ y $\mathbf{1}\{X < \varepsilon\} \geq 0$, el primer término es mayor o igual a 0. Por la monotonía de la esperanza:

$$EX \geq \varepsilon\,E\big(\mathbf{1}\{X \geq \varepsilon\}\big) = \varepsilon\,P(X > \varepsilon). \qquad \square$$

Corolario 13. *Si* $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ *es no decreciente, entonces*

$$P(X \geq \varepsilon) \leq P(\varphi(X) \geq \varphi(\varepsilon)) \leq \frac{E\varphi(X)}{\varphi(\varepsilon)}$$

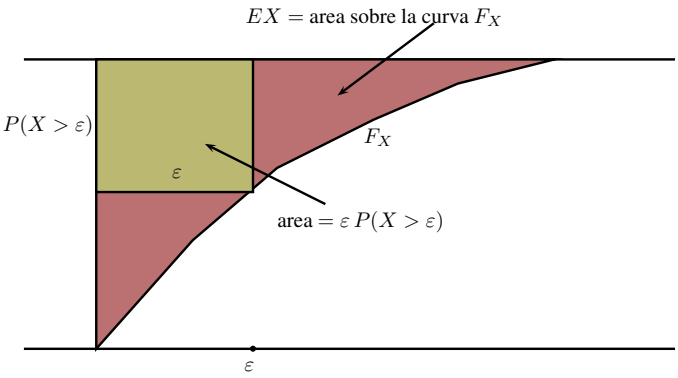


Figura 1: Desigualdad de Markov

Desigualdad de Chevichev

$$P(|X - EX| \geq \varepsilon) \leq \frac{VX}{\varepsilon^2}$$

Dem Usamos $\varphi(x) = x^2$ para obtener

$$P(|X - EX| \geq \varepsilon) \leq P\big(|X - EX|^2 \geq \varepsilon^2\big) \leq \frac{E(X - EX)^2}{\varepsilon^2}.$$

por la desigualdad de Markov.

La cota que provee la desigualdad de Chebyshev puede ser grosera o, peor aún, no informativa, por ejemplo, si $\varepsilon^2 \leq \sigma^2$

Ejemplo: Sea $X \sim U(0, 10)$, entonces $E(X) = 5$ y $V(X) = 100/12$.

Aplicando la desigualdad de Chebyshev,

$$P(|X - 5| > 4) \leq 0{,}52$$

Verdadero valor:

$$P(|X - 5| > 4) = 0{,}20$$

Convergencia en probabilidad: Sea X_n, $n \geq 1$, una sucesión de variables aleatorias, diremos que X_n converge en probabilidad a la v.a. X si para todo $\varepsilon > 0$

$$\lim_n P(|X_n - X| > \varepsilon) = 0$$

Ley de grandes números:

Sea X una variable aleatoria con $EX = \mu$. Se desea estimar μ por $\bar{X}_n$, la media muestral de una muestra de X.

Teorema. *Sean $X_1, X_2, \ldots$ iid. $EX = \mu$ $VX = \sigma^2$. Entonces $\bar{X}_n$ converge a μ en probabilidad.*

Dem: Ya vimos que $E\bar{X}_n = \mu$, $V\bar{X}_n = \sigma^2/n$.

Chevichev:

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \square$$

Versión Bernoulli de la Ley de los Grandes Números:

Consideremos n repeticiones independientes de un experimento aleatorio y sea A un evento con probabilidad $P(A) = p$, constante en las n repeticiones. Si llamamos $\hat{p}_n$ la proporcion muestral de A (número de veces que ocurre A en las n repeticiones dividido n), entonces $\hat{p}_n$ converge en probabilidad a p.

Dem: Note que $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$, donde $X_i = 1$ si A ocurre en el i-esimo ensayo y $X_i = 0$ si no ocurre.

$X_i \sim X \sim$ Bernoulli p.

$EX = p$,$VX = p(1-p)$.

$$\bar{X}_n = \hat{p}_n$$

y se obtiene:

$$P\Big(\Big|\hat{p}_n - p\Big| > \varepsilon\Big) \leq \frac{p(1-p)}{n\varepsilon^2} \to 0, \text{ con } n.$$

Ejemplo: Cuántas repeticiones del experimento deberían hacerse para que la proporción muestral $\hat{p}_n$ difiera de p en menos de 0,01 con probabilidad mayor o igual que 0,95? En este caso, $\varepsilon = 0{,}01$ y queremos encontrar n tal que

$$P(|\hat{p}_n - p| < 0{,}01) \geq 0{,}95$$

que equivale a

$$P(|\hat{p}_n - p| \geq 0{,}01) \leq 0{,}05$$

Chevichev: $0{,}05 = p(1-p)/(0{,}01^2 n)$ y se despeja n:

$$n \geq \frac{p(1-p)100^2}{0{,}05^2}$$

Tomando el mayor valor posible de $p(1-p) \leq \frac{1}{4}$, es suficiente tomar

$$n \geq \frac{1}{4}10^4\frac{10^4}{25} = \frac{10^8}{10^2} = 10^6.$$

10. Función generadora de momentos

Definición: momento de orden k de X, EX^k siempre que la esperanza exista.

$E(X) = \mu$ 1er momento: posición

$E(X^2) = \sigma^2 + \mu^2$ 2do momento: medida de dispersión

$E(X^3)$ 3er momento: medida de asimetría

$E(X^4)$ 4to momento: kurtosis (puntiaguda o chata)

Definición La *función generadora de momentos* de X, denotada $M_X : \mathbb{R} \to \mathbb{R}$, está definida por

$$M_X(t) = E(e^{tX})$$

si existe para $t \in (-h, h)$ para algún h. Esta es una condicion técnica para que $M(t)$ sea diferenciable en 0.

Los momentos determinan la FGM

Desarrollando en serie $e^\lambda = \sum_{k=0}^{\infty} \lambda^k/k!$, obtenemos

$$M_X(t) \;=\; E(e^{tX}) = \sum_{k=0}^{\infty} EX^k \frac{t^k}{k!}$$

El intercambio de suma con esperanza: ????.

Porque generadora de momentos?

Teorema 14. *Sea X con FGM $M_X(t)$. Entonces*

$$EX^n = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

Dem. Prueba corta:

$$\frac{d^n}{dt^n} M_X(t) = E\Big(\frac{d^n}{dt^n} e^{tX}\Big) = E(X^n e^{tX})$$

(Pero hay que justificar el pase de la derivada dentro de la esperanza.) Calculando en $t = 0$ obtenemos el teorema.

Para entender mejor tratamos los casos discreto y continuo

$$\frac{d^n}{dt^n} M_X(t) = \frac{d^n}{dt^n} E e^{tX} = \frac{d^n}{dt^n} \sum_x e^{tx} p(x) = \sum_x \frac{d^n}{dt^n} e^{tx} p(x)$$

$$= \sum_x x^n e^{tx} p(x)$$

que da EX^n al calcular la suma para $t = 0$.

La misma cuenta vale para el continuo:

$$\frac{d^n}{dt^n} M_X(t) = \frac{d^n}{dt^n} \int_{-\infty}^{\infty} e^{tx} f(x) dx = \int_{-\infty}^{\infty} \frac{d^n}{dt^n} e^{tx} f(x) dx$$

$$= \int_{-\infty}^{\infty} x^n e^{tx} f(x) dx.$$

que da EX^n al calcular la integral para $t = 0$. □

Ejemplos

Exponencial

$$M(t) = \frac{\lambda}{\lambda - t}$$

Momentos

$$EX = \frac{1}{\lambda} \qquad VX = \frac{1}{\lambda^2}$$

binomial

$$M(t) = (e^{tp} + (1 - p))^n$$

media varianza

Propiedad $Y = aX + b$ entonces

$$M_Y(t) = e^{bt} M_X(at)$$

Teorema de Unicidad: *Si existe la función generadora de momentos de una variable aleatoria, es única. Además la función generadora de momentos determina a la función de densidad o probabilidad de la v.a. salvo a lo sumo en un conjunto de probabilidad 0.*

Vamos a probar el teorema solo cuando X asume un número finito de enteros no negativos.

Prueba cuando $R_X = \{0, \ldots, n\}$ Fije $p(j) = P(X = j)$ y escriba

$$M(t) = \sum_{j=1}^{n} e^{tj} p(j)$$

$M(t)$ es un polinomio en $z = e^t$. Si definimos

$$H(z) = \sum_{j=0}^{n} z^j p(j)$$

H es llamada la *función generatriz*. H es un polinomio en z que da la misma info que M. Conocemos H si y solo si conocemos M. H es un polinomio de grado n.

Derivando j veces en z, tenemos que:

$$\left[H^{(j)}(z) \right]_{z=0} = j! p(j) \tag{23}$$

De donde,

$$p(j) = [\text{coeficiente de } z^j \text{ en } H(z)] = \frac{H^{(j)}(0)}{j!} \qquad \square$$

Ejemplo Sea X con momentos $\mu_k = EX^k$ dados por

$$\mu_0 = 1, \qquad \mu_k = \frac{1}{2} + \frac{2^k}{4}, \qquad \text{para } k \geq 1$$

Calcule la distribución de X.

$$M(t) = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!} = 1 + \frac{1}{2}\sum_{k=1}^{\infty} \frac{t^k}{k!} + \frac{1}{4}\sum_{k=1}^{\infty} \frac{(2t)^k}{k!} = \frac{1}{4} + \frac{1}{2}e^t + \frac{1}{4}e^{2t}$$

$$H(z) = \frac{1}{4} + \frac{1}{2}z + \frac{1}{4}z^2$$

De donde $p(0) = \frac{1}{4}$, $p(1) = \frac{1}{2}$, $p(2) = \frac{1}{4}$. Es decir, X tiene distribución Binomial $(2, 1/2)$.

FGM de Sumas de variables independientes

Teorema *Si X_i son variables aleatorias independientes con FGM $M_i(t)$ entonces:*

$$M_{X_1+\cdots+X_n}(t) = M_1(t) \ldots M_n(t)$$

Dem. Por independencia,

$$M_{X_1+\cdots+X_n}(t) = E(e^{t(X_1+\cdots+X_n)}) = E(e^{tX_1} \ldots e^{tX_n})$$
$$= Ee^{tX_1} \ldots Ee^{tX_n} = M_1(t) \ldots M_n(t). \qquad \square$$

Otras propiedades:

1) $M_{aX+b}(t) = e^{tb}Ee^{atX} = e^{tb}M_X(at)$

2) Si $Z \sim N(0,1)$, entonces $M_Z(t) = e^{t^2/2}$:

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx$$

$$= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}(x^2 - 2xt + t^2)\Big) dx = e^{t^2/2}$$

3) Si $X \sim N(\mu, \sigma^2)$, entonces $M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} e^{\sigma^2 t^2/2}$

4) Si $X_1, \ldots, X_n$ son iid media μ varianza σ^2 y $S_n = X_1 + \cdots + X_n$,

$$M_{S_n} = (M_X(t))^n$$

5) Si $T_n = S_n/\sqrt{n}$,

$$M_{T_n} = (M_X(t/\sqrt{n}))^n$$

6) Suma de normales independientes es normal con media = suma de las medias y varianza igual a la suma de las varianzas.

11. Teorema central del límite

Convergencia en distribución: Decimos que una sucesión de variables aleatorias $Y_1, Y_2, \ldots$ converge en distribución a una variable Y si

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

para todo y donde $F_Y(y)$ es continua.

Teorema de Unicidad de la FGM. *Si la FGM de una variable aleatoria existe, entonces es única. Además la FGM de X determina la función de distribución acumulada F_X.*

Convergencia en distribución es equivalente a convergencia de las FGM:

$$Y_n \to Y \text{ en distribución } \underline{\text{sii}} \lim_{n\to\infty} M_{X_n}(t) = M_X(t).$$

Teorema central del limite. *Sean X_i iid con media μ y varianza σ^2 y sea $S_n := X_1 + \cdots + X_n$. Entonces*

$$Z_n := \frac{S_n - n\mu}{\sigma\sqrt{n}} \longrightarrow Z, \text{ en distribución,}$$

donde $Z \sim N(0,1)$.

Observaciones:

1) Z_n tiene media 0 y varianza 1 para todo n.

2) Convergencia en distribución es Convergencia de las acumuladas.

3) Uso: para n grande trate Z_n como si fuera $N(0,1)$.

Historia:

1733: TCL para Bernoulli(1/2) por Abraham de Moivre

1823: Pierre-Simon Laplace extiende de Moivre's para aproximar la Binomial(n,p) por la normal.

1901: Aleksandr Lyapunov demuestra rigurosamente el TCL.

Demostración del TCL:

Asumimos que la FGM $M = M_{X_i}$ de X_i existe e inicialmente tomamos $\mu = 0$ y $\sigma^2 = 1$ (despues vemos como se extiende).

Calculamos la FGM de Z_n en función de M:

$$M_{Z_n}(t) = Ee^{t(X_1+\cdots+X_n)/\sqrt{n}} = (M(t/\sqrt{n}))^n$$

Sea

$$L(t) = \log M(t)$$

y note que

$$L(0) = 0, \quad L'(0) = \frac{M'(0)}{M(0)} = \mu = 0$$

$$L''(0) = \frac{M(0)M''(0) - (M'(0))^2}{(M(0))^2}$$

$$= \frac{EX^2 - (EX)^2}{1} = EX^2 = 1$$

Para probar el teorema, necesitamos probar que

$$\lim_{n\to\infty} (M(t/\sqrt{n}))^n = e^{t^2/2}$$

que es equivalente a probar que

$$\lim_{n\to\infty} nL(t/\sqrt{n}) = t^2/2.$$

Calculemos:

$$\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{n^{-1}} = \lim_{n\to\infty} \frac{L'(t/\sqrt{n})tn^{-3/2}}{2n^{-2}}$$

(por L'Hopital)

$$= \lim_{n\to\infty} \frac{L'(t/\sqrt{n})t}{2n^{-1/2}} = \lim_{n\to\infty} \frac{L''(t/\sqrt{n})t^2n^{-3/2}}{2n^{-3/2}}$$

(de nuevo por L'Hopital)

$$= \lim_{n\to\infty} L''(t/\sqrt{n})\frac{t^2}{2} = \frac{t^2}{2}.$$

Esto termina la demostración para media cero y varianza 1.

Si μ y σ^2 son cualesquiera,

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}}\Big(\frac{X_1 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma}\Big)$$

y se aplica la demostración anterior a las variables $X_i^* = \frac{X_i - \mu}{\sigma}$ que son centradas y tienen varianza 1. $\square$

Formas alternativas del TCL:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to Z$$

y dividiendo numerador y denominador por n, obtenemos

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to Z$$

Una razón matemática para el TCL:

$$Z_{2n} = \frac{S_{2n}}{\sqrt{2n}} = \frac{S_n + S_{2n} - S_n}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\Big(\frac{S_n}{\sqrt{n}} + \frac{S_n^*}{\sqrt{n}}\Big),$$

donde S_n^* tiene la misma distribución que S_n pero es independiente de S_n.

O sea que el límite, si existe tiene que satisfacer:

$$Z \sim \frac{Z + Z^*}{\sqrt{2}} \qquad (*)$$

para Z y Z^* identicamente distribuídas e independientes. En términos de la FGM esa ecuación es equivalente a

$$M_Z(t) = (M_Z(t/\sqrt{2}))^2$$

que es satisfecha por la normal:

$$M_Z(t) = e^{t^2/2} = (e^{(t/\sqrt{2})^2/2})^2 = (M_Z(t/\sqrt{2}))^2$$

Para obtener una demostración del TCL usando este argumento falta probar: (1) que el limite de Z_n existe y (2) que la normal es la *única* distribución que satisface la "ecuación" (*).

Comentarios sobre el TCL. Qué significa n suficientemente grande? Cómo sabemos si la aproximación es buena? El tamaño de muestra requerido para que la aproximación sea razonable depende de la forma de la distribución de las X_i. Mientras más simétrica y acampanada sea, más rápidamente se obtiene una buena aproximación.

Ejemplo: Al sumar números, una calculadora aproxima cada número al entero más próximo. Los errores de aproximación se suponen independientes y con distribución U(-0.5,0.5).

a) Si se suman 1500 números, ¿cuál es la probabilidad de que el valor absoluto del error total exceda 15?

Si llamamos X_i al error correspondiente al i-ésimo sumando, el error total es $T_{1500} = \sum_i X_i$ y queremos calcular $P(|T_{1500}| > 15)$. Como $EX_i = 0$ y $VX_i = 1/12$, $ET_{1500} = 0$ y $VT_{1500} = \frac{1500}{12} = 125$. Entonces

$$P(|T_{1500}| > 15) = P(|Z| > 15/\sqrt{125}) = P(|Z| > 1{,}34) = 0{,}18$$

(usando la tabla de la Normal)

b) ¿Cuántos números pueden sumarse a fin de que el valor absoluto del error total sea menor o igual que 10 con probabilidad mayor o igual que 0.90? Buscamos el valor de n tal que $P(|T_n| \leq 10) \geq 0{,}9$.

$$P(|T_n| \leq 10) \geq 0{,}9 \Leftrightarrow P(|Z| \leq 10/\sqrt{n/12}) \geq 0{,}9$$

Buscamos z tal que $P(|Z| \leq z) = 0{,}9$, que por tabla es $z = 1{,}64$. Así

$$10/\sqrt{n/12} = 1{,}64, \text{ de donde } n \geq 446.$$

Otras Aplicaciones del TCL

1. Si $Y_n \sim$ Poisson (λn) entonces

$$\frac{Y_n - n\lambda}{\sqrt{n}\lambda} \xrightarrow{D} Z$$

Dem: considere X_i Poisson(λ) iid.

$Y_n = X_1 + \cdots + X_n$ Poisson (λn). Aplique TCL y obtenga el límite.

Así la Poisson con parametro grande se aproxima por la normal.

2. $Y_n \sim$ Gama(n, λ) iid con n entero

$$\frac{Y_n - n\lambda}{\sqrt{n}\lambda} \xrightarrow{D} Z$$

$X_i \sim$ Gama$(1, \lambda)$ (exponenciales) independientes.

$X_1 + \cdots + X_n$ Gama (n, λ) suma de n exponenciales independientes.

Así la suma de gamas se aproxima por la normal.

3. Un adivino acierta el color de 950 de 1500 cartas puestas al dorso. Queremos decidir si creemos que es adivino.

Sea p la probabilidad que el adivino acierte. Queremos testar $p = 1/2$ (es decir, no mejora el puro azar) contra $p > 1/2$ (tiene probabilidad de adivinar mayor que $1/2$).

Supongamos que decide al azar, $p = 1/2$.

Sea $X_i = \mathbf{1}\{\text{acierta la carta } i\}$. Azar $\Rightarrow X_i \sim$ Bernoulli$(\frac{1}{2})$

Número de aciertos:

$$S_{1500} = \sum_{i=1}^{1500} X_i, \qquad \bar{X} = \frac{S_{1500}}{1500}$$

$$P(S_{1500} \geq 950) = P\left(\frac{\bar{X} - \frac{1}{2}}{0,5/\sqrt{1500}}\right) \geq \frac{950/1500}{0,5/\sqrt{1500}}$$

$$\sim P(Z \geq 10,32) \sim 0$$

La proba de acertar 950 veces con una moneda es casi 0. Aceptamos la hipótesis que el hombre es un adivino.

Porqué convergencia en puntos de continuidad de F?

Considere una sucesión de variables aleatorias X_n con acumuladas $F_n(x) = \mathbf{1}\{x \geq 1/n\}$.

X_n es una variable aleatoria constante: $P(X_n = 1/n) = 1$.

Cuando $n \to \infty$, la distribución de X_n aproxima la distribución de una variable aleatoria X concentrada en 0: $P(X = 0) = 1$. Sin embargo, si F es la acumulada de X, vemos que $F_n(0)$ no converge a $F(0)$.

De hecho, $F_n(0) = 0$ para todo n, pero $F(0) = 1$.

12. Procesos de Bernoulli y Poisson

12.1. Procesos de Bernoulli y Binomial

Un **proceso estocástico** es una sucesión de variables aleatorias $X_1, X_2, \ldots$ indexadas por $n \in \mathbb{N}$ o $t \in \mathbb{R}$.

Si X_i son iid Bernoulli(p), diremos que el proceso $X_1, X_2, \ldots$ es un **Proceso de Bernoulli**. Se trata de una sucesión de variables aleatorias independientes Bernoulli(p).

El proceso de Bernoulli es **estacionario**:

$$P(X_1 = b_1, \ldots, X_n = b_n) = P(X_{t+1} = b_1, \ldots, X_{t+n} = b_n)$$

para todo t.

Ejemplo: En la parada del pabellón 2 sale un **colectivo 107** en cada minuto con probabilidad 1/10, en forma independiente. Cual es la probabilidad que salgan colectivos en los minutos 1,2,3? Y en los minutos

27,28,29?

$$P(X_1 = 1, X_2 = 1, X_3 = 1) = \left(\frac{1}{10}\right)^3.$$

$$P(X_{27} = 1, X_{28} = 1, X_{29} = 1) = \left(\frac{1}{10}\right)^3.$$

Proceso Binomial

Definamos las variables $S_n = X_1 + \cdots + X_n$. El proceso

$$S_1, S_2, \ldots$$

es llamado *proceso Binomial*. S_n cuenta el número de éxitos hasta el n-ésimo ensayo. S_n tiene distribución Binomial(n, p) para cada $n \geq 1$.

El **incremento** del proceso en el intervalo $(m, n]$, con $n \leq m$ está dado por

$$S_n - S_m = \sum_{k=m+1}^{n} X_k.$$

El proceso binomial tiene **incrementos estacionarios**:

$$P(S_{n+m} - S_m = k) = P(S_n = k) \tag{24}$$

La distribución del número de llegadas en un intervalo depende del tamaño del intervalo y no de su localización.

El proceso binomial tiene **incrementos independientes**: Si $1 \leq m \leq n \leq i \leq j$,

$$\begin{aligned} P(S_n - S_m = k, S_j - S_i = h) &= P(S_n - S_m = k)P(S_j - S_i = h) \\ &= P(S_{m-n} = k)P(S_{i-j} = h) \end{aligned}$$

La probabilidad de incrementos k y h en los intervalos disjuntos $(m, n]$ y $(i, j]$, respectivamente, es el producto de las probabilidades. Más generalmente, vale para conjuntos finitos de intervalos:

$$\begin{aligned} &P(S_{n_1} - S_{m_1} = k_1, \ldots, S_{n_\ell} - S_{m_\ell} = k_\ell) \\ &\quad = P(S_{n_1} - S_{m_1} = k_1) \ldots P(S_{n_\ell} - S_{m_\ell} = k_\ell) \\ &\quad = P(S_{n_1 - m_1} = k_1) \ldots P(S_{n_\ell - m_\ell} = k_\ell) \end{aligned} \tag{25}$$

si los intervalos $[m_j, n_j]$ son disjuntos dos a dos.

Instante de la primera llegada $Y_1 := \text{mín}\{k > 0 : X_k = 1\}$ tiene distribución geométrica:

$$P(Y_1 = k) = P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) = (1 - p)^{k-1}p$$

El evento $\{Y_1 = k\}$ depende de $X_1, \ldots, X_k$, por lo tanto se puede calcular su probabilidad.

Ejemplo: Colectivo Si llego en un instante t cualquiera y defino el tiempo de espera del colectivo a partir de ese instante:

$$R_t := \text{mín}\{k > 0 : X_{t+k} = 1\}$$

$$P(R_t = k) = P(X_{t+1} = 0, \ldots, X_{t+k-1} = 0, X_{t+k} = 1) = (1 - p)^{k-1}p$$

Tiene distribución geométrica igual que si empezaba en el instante 0.

Instante de la k-ésima llegada

$$Y_k := \text{mín}\{n : X_1 + \cdots + X_n = k\}$$

Para $t \geq k$:

$$\begin{aligned} P(Y_k = t) &= P(k-1 \text{ exitos en } [1, t-1], \text{éxito en } t) \\ &= \binom{t-1}{k-1} p^{k-1}(1-p)^{t-1-(k-1)} p \end{aligned}$$

Es decir que el instante de la k-ésima llegada tiene distribución Binomial negativa de parámetros k y p.

Dualidad entre Binomial y Binomial negativa

La k-ésima llegada ocurre antes del instante n si y sólo si el número de llegadas hasta el instante n es mayor o igual a k:

$$Y_k \leq n \quad \Longleftrightarrow \quad S_n \geq k,$$

donde $S_n = X_1 + \cdots + X_n$ y $Y_k := \text{mín}\{n : X_1 + \cdots + X_n = k\}$, tienen distribución $S_n \sim \text{Binomial}(n, p)$ y $Y_k \sim$ Binomial negativa (k, p).

Tiempo entre llegadas sucesivas

Sea $T_0 := 0$ y $T_i := Y_i - Y_{i-1}$, $i \geq 1$. Ya vimos que Y_i tiene distribución binomial negativa.

Lema 15. *$(T_k,\ k \geq 1)$ son variables independientes con distribución geométrica (p).*

Dem Escribiendo $s_j = t_1 + \cdots + t_j$ tenemos

$$\begin{aligned} P(T_1 = t_1, \ldots, T_k = t_k) &= P(\cap_{j=1}^{k} \{X_{s_{j-1}+1} = \cdots = X_{s_j - 1} = 0, X_{a_j} = 1\}) \\ &= (1-p)^{t_1 - 1} p \ldots (1-p)^{t_k - 1} p \end{aligned}$$

Sumando esa expresión sobre todos los $t_k \geq 1$ para $k \neq i$, obtenemos que la i-ésima marginal tiene distribución geométrica(p): $P(T_i = t_i) = (1-p)^{t_i - 1} p$. Concluimos que

$$P(T_1 = t_1, \ldots, T_k = t_k) = P(T_1 = t_1) \ldots P(T_k = t_k),$$

lo que prueba la independencia. □

Como corolario tenemos que

$$P(T_1 > t_1, \ldots, T_k > t_k) = P(T_1 > t_1) \ldots P(T_k > t_k)$$

Resumen de procesos

$(X_1, X_2, \ldots)$ Bernoulli(p) independientes

$(S_1, S_2, \ldots)$ donde $S_n \sim \text{Binomial}(n, p)$ con incrementos independientes y estacionarios.

$(T_1, T_2, \ldots)$, Geometricas(p) independientes.

$(Y_1, Y_2, \ldots)$, donde $Y_j \sim$ Binomial negativa (j, p).

12.2. Proceso de Poisson

Un proceso puntual es un subconjunto discreto $S \subset \mathbb{R}^+$. Llamaremos llegadas a los puntos del proceso. Llamamos $N_t = \big|S \cap [0, t]\big|$, el número de llegadas de S en el intervalo $[0, t]$. Sea $\lambda > 0$ un parámetro.

Definición Decimos que $(N_t : t \in \mathbb{R}^+)$ es un *proceso de Poisson* de intensidad λ si su distribución satisface:

i) N_t tiene distribución Poisson(λt) para todo t.

ii) *Incrementos estacionarios.* El número de llegadas en un intervalo depende sólo del tamaño del intervalo: $N_{a+t} - N_a$ tiene la misma distribución que N_t, para todo $a > 0, t > 0$.

iii) *Incrementos independientes.* Las llegadas en intervalos disjuntos son independientes: Si $(s_i, t_i]$, $i = 1, \ldots, k$, son intervalos disjuntos dos a dos, entonces las variables aleatorias $(N_{t_i} - N_{s_i})$, $i = 1, \ldots, k$, son independientes.

Como los incrementos son estacionarios, el número de llegadas en cualquier intervalo $(s, t]$ tiene distribución Poisson$(\lambda(t-s))$.

Ejemplo Los mails llegan a una computadora de acuerdo a un proceso de Poisson de intensidad $\lambda = 2$ mensajes/minuto. Sea N_t = número de mensajes entre 0 y t.

a) ¿Cuál es la probabilidad de que no se reciba ningún mensaje entre las 12 y las 12:03? $N_3 \sim$ Poisson(2·3) =Poisson(6). $P(N_3 = 0) = e^{-6} = 0{,}002$.

b) ¿Cuál es la probabilidad de que no se reciba ningún mensaje entre las 13:30 hs y las 13:33 hs? Misma respuesta que en (a).

Construcción usando exponenciales independientes

Sean $\tau_1, \tau_2, \ldots$ iid Exponencial(λ). Sea $Y_0 = 0$ y $Y_n := \tau_1 + \cdots + \tau_n$. Defina

$$M(s) := \text{máx}\{n : Y_n \leq s\}$$

τ_n tiempos entre llegadas a un banco.

Y_n = instante de la n-ésima llegada. $Y_n \sim \text{Gama}(n, \lambda)$.

$M(t)$ es el número de llegadas hasta el instante t. Tenemos

$$M(s) \geq n \text{ si y sólo si } Y_n \leq s$$

Por lo tanto, la dualidad Gama-Poisson implica

$$M(s) \sim \text{Poisson}(\lambda s)$$

Lema 16. *El proceso* $(\tilde{M}(t) : t \geq 0) := (M(t+s) - M(s) : t \geq 0)$ *tiene la misma distribución que* $(M(t) : t \geq 0)$*, para cada* $s \geq 0$*.*

Demostración. Si llamamos $\tilde{\tau}_1, \tilde{\tau}_2, \ldots$ los tiempos entre llegadas después de s, condicionando a $\{M(s) = n, Y_n = u\}$ con $u \in [0, s)$, tendremos $\tilde{\tau}_1 = Y_{n+1} - s$ y su distribución se puede calcular así:

$$\begin{aligned} P(\tilde{\tau}_1 > t | Y_n = u, M(s) = n) &= P(Y_{n+1} - s > t | Y_n = u, M(s) = n) \\ &= P(Y_{n+1} - s > t | Y_n = u, Y_{n+1} > s) \\ &= P(\tau_{n+1} > t + s - u | \tau_{n+1} > s - u, Y_n = u) \\ &= P(\tau_{n+1} > t + s - u | \tau_{n+1} > s - u) = e^{-\lambda t} \end{aligned}$$

donde la segunda y tercera identidades son consecuencia de la identidad entre los eventos condicionantes; la tercera se deduce de la independencia entre Y_n y τ_{n+1} y la tercera es la falta de memoria de la exponencial. Como, por probabilidad total tenemos

$$\begin{aligned} &P(\tilde{\tau}_1 > t) \\ &= \sum_{n \geq 0} \int_0^s P(Y_{n+1} - s > t | Y_n = u, M(s) = n) P(M(s) = n) f_{Y_n | M(s) = n}(u) du \\ &= e^{-\lambda t} \sum_{n \geq 0} \int_0^s P(M(s) = n) f_{Y_n | M(s) = n}(u) du = e^{-\lambda t}. \end{aligned}$$

podemos concluir que $\tilde{\tau}_1 \sim \text{Exponencial}(\lambda)$.

Para los intervalos entre llegadas sucesivos se hace el mismo cálculo: condicionando a $\{Y_n = u, M(s) = n\}$ tenemos $\tilde{\tau}_1 = Y_{n+1} - s = \tau_{n+1} - (s - u)$ y $\tilde{\tau}_j = \tau_{n+j}$ para $j \geq 2$. Por ejemplo:

$$\begin{aligned} &P(\tilde{\tau}_2 > t_2, \tilde{\tau}_1 > t_1 \,|\, Y_n = u, M(s) = n) \\ &\quad = \cdots = P(\tau_{n+1} > t_1 + s - u, \tau_{n+2} > t_2 | \tau_{n+1} > s - u, Y_n = u) \\ &\quad = e^{-\lambda t_1} e^{-\lambda t_2}. \end{aligned}$$

Para ver que $\tilde{\tau}_1, \tilde{\tau}_2, \ldots$ es independiente de $(M(t), t \leq s)$, basta ver que el cálculo de la distribución de $\tilde{\tau}_j$ depende de $(M(t), t \leq s)$ solamente a traves de $Y_{M(s)}$ y vimos que en realidad es independiente de $Y_{M(s)}$. □

Lema 17. *El proceso* $M(t)$ *tiene incrementos independientes: si* $t_0 < t_1 < \ldots < t_n$ *entonces*

$$M(t_1) - M(t_0), \ldots, M(t_n) - M(t_{n-1}) \text{ son independientes.}$$

Demostración. El Lema 16 dice que $M(t_n) - M(t_{n-1})$ es independiente de $M(r), r \leq t_{n-1}$ y por lo tanto de $M(t_1) - M(t_0), \ldots, M(t_{n-1}) - M(t_{n-2})$. Concluya usando inducción. □

Corolario 18. *El proceso* $M(t)$ *construido con las exponenciales es un proceso de Poisson.*

12.3. El Proceso Binomial aproxima al Proceso de Poisson

Fijamos un $\lambda > 0$ y consideramos una sucesión de procesos de Bernoulli, indexados por $\ell > 0$, $(X_n^\ell, n \in \mathbb{N})$, cada uno consiste en variables iid Bernoulli(λ/ℓ). Es decir, $P(X_n^\ell = 1) = p(\ell) = \lambda/\ell$.

Vamos a introducir una familia de procesos Binomiales indexados por ℓ donde los ensayos ocurren a cada $1/\ell$ instantes y la probabilidad de éxito en cada ensayo es λ/ℓ. Sea t real positivo y defina el proceso $(S_t^\ell, t \in \mathbb{R}^+)$ por

$$S_t^\ell := \sum_{n:(n/\ell)\le t} X_n^\ell,$$

el número de éxitos hasta el instante t. Son $[\ell t]$ ensayos de Bernoulli independientes, cada uno con probabilidad λ/ℓ de éxito. S_t^ℓ es un proceso Binomial definido en la grilla $\mathbb{N}/\ell$. El número esperado de éxitos en el intervalo $[0, t]$ es

$$ES_t^\ell = \frac{\lambda}{\ell}[t\ell] = \lambda t + O(1/\ell).$$

Vimos en (24)-(25) que para cada ℓ, el proceso binomial S_t^ℓ tiene incrementos estacionarios e independientes.

Teorema 19. *Cuando $\ell \to \infty$, las distribuciones finito-dimensionales de los procesos binomiales $(S_t^\ell, t \in \mathbb{R}^+)$ convergen a las distribuciones finito-dimensionales del proceso de Poisson $(N_t, t \in \mathbb{R}^+)$.*

Dem Veremos que el límite satisface i-iii de la definición del Proceso de Poisson.

i) Las variables S_t^ℓ convergen en distribución a variables Poisson(λ),

$$\lim_\ell P(S_t^\ell = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} = P(N_t = k), \qquad k \in \{0, 1, \dots\} \tag{26}$$

por la aproximación de la Binomial a la Poisson.

ii-iii) Por los incrementos independientes del proceso binomial (25), tenemos que si $(s_i, t_i]$, $i = 1, \dots, k$, son intervalos temporales disjuntos dos a dos, entonces las variables aleatorias $(S_{t_i}^\ell - S_{s_i}^\ell)$, $i = 1, \dots, k$, son independientes:

$$P(S_{t_i}^\ell - S_{s_i}^\ell = h_i,\ i = 1, \dots, k) = \prod_{i=1}^k P(S_{t_i}^\ell - S_{s_i}^\ell = h_i) \tag{27}$$

Sacando límite en ℓ, usando los incrementos estacionarios (24) y la convergencia a Poisson (26) obtenemos

$$\lim_{\ell\to\infty} P(S_{t_i}^\ell - S_{s_i}^\ell = h_i,\ i = 1, \dots, k) = \prod_{i=1}^k \frac{e^{-\lambda(t_i - s_i)}(\lambda(t_i - s_i))^{h_i}}{h_i!} \tag{28}$$

$$= P(N_{t_i} - N_{s_i} = h_i,\ i = 1, \dots, k) \tag{29}$$

Es decir que el vector $(S_{t_i}^\ell - S_{s_i}^\ell,\ i = 1, \dots, k)$ converge en distribución al vector $(N_{t_i} - N_{s_i},\ i = 1, \dots, k)$. □

Convergencia de los tiempos entre llegadas

Sea $Y_n^\ell := \min\{t > 0 : S_t^\ell = n\}$ el tiempo de la n-ésima llegada en S_t^ℓ. Como $\ell Y_1^\ell \sim$ Geometrica(λ/ℓ),

$$P(Y_1 > t) = \lim_{n\to\infty} P(\ell Y_1^\ell > \ell t) = \lim_{\ell\to\infty} \Big(1 - \frac{\lambda}{\ell}\Big)^{[\ell t]} = e^{-\lambda t}$$

Sea $T_i^\ell := Y_i^\ell - Y_{i-1}^\ell$ el tiempo entre la $(i-1)$-ésima llegada y la i-ésima llegada del proceso Binomial S_t^ℓ, con la convención $Y_0 = 0$.

Lema 20. *Los tiempos entre llegadas $(T_i^\ell : i \ge 0)$ del proceso binomial S_t^ℓ convergen en distribución a iid Exponencial(λ).*

Dem Si ℓt_i es entero (si no es entero cometemos un error de orden $1/\ell$ que se va a cero) vimos en el Lema 15 que

$$P(T_i^\ell > t_i, i = 1, \ldots, k) = \prod_{i=1}^{k} (1 - \lambda/\ell)^{\ell t_i - 1} \lambda/\ell,$$

es decir que T_i^ℓ tienen la misma distribución que k variables independientes geométricas de parámetro λ/ℓ, divididas por ℓ. Usando que las probabilidades puntuales caracterizan las probabilidades acumuladas, obtenemos en particular la convergencia en distribución de los primeros k tiempos entre llegadas:

$$\begin{aligned} P(T_i > t_i, 1 \le i \le k) &= \lim_{\ell \to \infty} P(T_i^\ell > t_i, 1 \le i \le k) = \lim_{\ell \to \infty} \prod_{i=1}^{k} P(T_i^\ell > t_i) \\ &= \lim_{\ell \to \infty} \prod_{i=1}^{k} (1 - \lambda/\ell)^{\ell t_i} = \prod_{i=1}^{k} e^{-\lambda t_i}. \qquad \square \end{aligned}$$

Distribución de un número fijo de puntos

Supongamos que hay una única llegada en $[0, t]$. Cual es la distribución del instante de esa llegada? Sea $s \in [0, t]$ y calcule

$$P\Big(\frac{T_1^\ell}{\ell} \le s \Big| S_t^\ell = 1\Big) = \frac{P(S_s^\ell = 1) P(S_t^\ell - S_s^\ell = 0)}{P(S_t^\ell = 1)} \to s$$

Sabiendo que hay k llegadas en el intervalo $[0, t]$, cual es la distribución de los instantes de llegada?

Teorema *En un proceso de Poisson*

$$P(\{Y_1, \ldots, Y_k\} \in B | N_t = k) = P(\{U_1, \ldots, U_k\} \in B)$$

donde $U_1, \ldots, U_k$ son variables aleatorias independientes Uniforme$[0, t]$.

Demostración. Hagamos el cálculo para $n = 3$. Estamos condicionando a $\{N(t) = 3\}$.

Basta demostrar que la densidad de (T_1, T_2, T_3) condicionada a $\{N(t) = 3\}$ es la misma que la de los estadísticos de orden de 3 Uniformes en $[0, t]$.

$$\begin{aligned} &\text{“}P(T_1 = t_1, T_2 = t_2, T_3 = t_3, \tau_4 > t - t_3 \,|\, N(t) = 3)\text{”} \\ &= \frac{\lambda e^{-\lambda t_1} \lambda e^{-\lambda(t_2 - t_1)} \lambda e^{-\lambda(t_3 - t_2)} e^{-\lambda(t - t_3)}}{e^{-\lambda t} (\lambda t)^3 / 3!} \mathbf{1}\{0 < t_1 < t_2 < t_3 < t\} \\ &= \frac{3!}{t^3} \mathbf{1}\{0 < t_1 < t_2 < t_3 < t\} \end{aligned}$$

O sea que la distribución de (T_1, T_2, T_3) condicionada a $N(t) = 3$ es uniforme en el conjunto $\{(t_1, t_2, t_3) : 0 < t_1 < t_2 < t_3 < t\}$. Ese conjunto tiene volumen $t^3/3!$ porque $\{(t_1, t_2, t_3) : 0 < t_1, t_2, t_3 < t\}$ tiene volumen t^3 y ese es uno de los 3! posibles órdenes entre las coordenadas.

El mismo razonamiento sirve para demostrar que la densidad de $(T_1, \ldots, T_n)$ dado $N(t) = n$ está dada por $n!/t^n$. $\square$ $\square$

Es decir que si $N(t) = n$, la posición de los puntos es la misma que la de n uniformes en $[0, t]$.

12.4. Construcción alternativa del proceso de Poisson

Elija un número aleatorio de puntos con distribución $N_T \sim \text{Poisson}(\lambda T)$. Distribuya ese número de puntos como iid uniformemente distribuídos en el intervalo $(0, T)$. Llámelos $U_1, \ldots, U_{N_T}$. Defina:

$$N_t := \sum_{i=1}^{N_T} \mathbf{1}\{i : U_i \le t\}$$

el número de puntos que cayeron a la izquierda de t.

Lema 21. *El proceso* $(N_t : 0 \leq t \leq T)$ *así construído es un proceso de Poisson en* $[0, T]$.

Extensión a dimensiones mayores Sea $l(B)$ la medida de Lebesgue del Boreliano B.

Un proceso de Poisson de intensidad λ en $\mathbb{R}^d$ es un subconjunto aleatorio S de $\mathbb{R}^d$ cuya distribución satisface

1. $|S \cap B| \sim \text{Poisson}(\lambda l(B))$, si B es un Boreliano.

2. $|S \cap B_1|, \ldots, |S \cap B_n|$ son independientes si los B_i son Borelianos disjuntos dos a dos.

La construcción 2 se puede extender a $\mathbb{R}^d$. Sea $\lambda > 0$ y considere

- Una partición $\mathcal{J}$ de $\mathbb{R}^d$ cuyos elementos son Borelianos acotados.
- Una familia $(Y_A : A \in \mathcal{J})$, donde Y_A son iid con $Y_A \sim \text{Poisson}(\lambda l(A))$.
- Una familia de sucesiones independientes $((U_{A,j}, j \geq 1), A \in \mathcal{J})$, donde $(U_{A,j} : j \geq 1)$ son iid con distribución uniforme en A.
- Defina el proceso de Poisson como el conjunto aleatorio dado por

$$S \stackrel{def}{=} \bigcup_{A \in \mathcal{J}} \bigcup_{j \leq Y_A} \{U_{A,j}\} = \bigcup_{A \in \mathcal{J}} \{U_{A,j} : j \leq Y_A\} \tag{30}$$

El objeto aleatorio así definido se llama *Proceso de Poisson* de *intensidad* λ.

13. Cadenas de Markov

13.1. Definición

Vamos a considerar procesos estocásticos a tiempo discreto $X_1, X_2, \ldots$, asumiendo valores en un conjunto S finito o numerable llamado *espacio de estados*. El sub-índice se interpreta como tiempo. Si $X_n = x$, diremos que el proceso se encuentra en el *estado* x en el *instante* n.

En una **cadena de Markov** cuando el proceso está en el estado x en el instante n, tiene probabilidad $Q(x, y)$ de ir al estado y en el instante $n + 1$, independientemente de la trayectoria que hizo para llegar a x:

$$P(X_{n+1} = y | X_n = x,\, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = Q(x, y).$$

Los valores $Q(x, y)$ son llamados *probabilidades de transición* y conforman una **matriz de transición** $Q = (Q(x, y) : x, y \in S)$. Esta matriz tiene entradas no negativas y la suma de cada fila vale 1:

$$\sum_{y \in S} Q(x, y) = 1$$

La matriz de transición es un conjunto de parámetros que caracterizan la cadena de Markov.

Cadena de Markov con dos estados. Si hoy llueve, la probabilidad que llueva mañana es 0,8 y si hoy no llueve, esta probabilidad es 0,1. El espacio de estados es $S = \{0, 1\}$. Si interpretamos 1 cuando llueve y 0 cuando no llueve, la matriz de transición es

$$Q = \begin{pmatrix} 0{,}9 & 0{,}1 \\ 0{,}2 & 0{,}8 \end{pmatrix} \tag{31}$$

$Q(0, 0) = 0{,}1$, $Q(0, 1) = 0{,}9$, $Q(1, 0) = 0{,}2$, $Q(1, 1) = 0{,}8$.

13.2. Construcción

Sea $U_1, U_2, \ldots$ una sucesión de variables Uniformes$[0, 1]$ independientes. Defina $X_0 = x \in \{0, 1\}$ e, iterativamente,

$$X_{n+1} = F(X_n, U_{n+1}) \tag{32}$$

donde $F(0,u) = \mathbf{1}\{u > 0{,}9\}$ y $F(1,u) = \mathbf{1}\{u > 0{,}2\}$.

Verifique que el proceso así obtenido es una cadena de Markov con matriz de transición (31).

Definición constructiva de cadenas de Markov

Sea Q una matriz de transición en un espacio de estados S. Para cada $x \in S$ definimos una partición $\mathcal{J}_x = (J(x,y), y \in S)$ del intervalo $[0,1]$, de tal manera que

$$|J(x,y)| = Q(x,y)$$

Defina $F : S \times [0,1] \to S$ por

$$F(x,u) = \sum_{y \in S} y\, \mathbf{1}\{u \in J(x,y)\}$$

Fije un estado x y defina un proceso estocástico X_n, $n \geq 0$ por $X_0 = x$ e iterativamente,

$$X_{n+1} = F(X_n, U_{n+1}) \tag{33}$$

El proceso así definido es Markov con matriz de transición Q. En efecto,

$$\begin{aligned} &P(X_{n+1} = y | X_n = x,\, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\ &\quad = P(F(x, U_{n+1}) = y | F(x_{n-1}, U_n) = x, \ldots, F(x_0, U_1) = x_1, X_0 = x_0)) \end{aligned}$$

Como los U_k son independientes, esa expresión es igual a

$$= P(F(x, U_{n+1}) = y) = P(U_{n+1} \in J(x,y)) = |J(x,y)| = Q(x,y).$$

13.3. Matriz de transición y distribución en el instante n

La matriz de transición sirve para calcular las probabilidades de transición a más de un paso:

$$P(X_n = y | X_0 = x) = Q^n(x,y) \tag{34}$$

Probemos esto para $n = 2$:

$$\begin{aligned} P(X_2 = y | X_0 = x) &= \sum_z P(X_2 = y, X_1 = z | X_0 = x) \\ &= \sum_z P(X_2 = y | X_1 = z, X_0 = x) P(X_1 = z | X_0 = x) \end{aligned}$$

(por las propiedades de probabilidad condicional)

$$= \sum_z P(X_2 = y | X_1 = z) P(X_1 = z | X_0 = x)$$

(por la propiedad de Markov)

$$= \sum_z Q(x,z) Q(z,y) = Q^2(x,y)$$

Para n general procedemos por inducción. Asuma (34) y calcule

$$P(X_{n+1} = y | X_0 = x) = \sum_z P(X_{n+1} = y, X_n = z | X_0 = x)$$

que por el mismo cálculo que antes es igual a

$$= \sum_z Q^n(x,z) Q(z,y) = Q^{n+1}(x,y)$$

Ecuaciones de Chapman-Kolmogorov

$$Q^{n+m}(x,y) = \sum_z Q^n(x,z) Q^m(z,y), \qquad 0 \leq k \leq n.$$

Demostración Por la fórmula (34),

$$Q^{n+m}(x,y) = P(X_{n+m} = y | X_0 = x) \tag{35}$$

Usando la partición $\{\{X_n = z\}, z\}$, por probabilidad total,

$$= \sum_z P(X_{n+m} = y, X_n = z | X_0 = x)$$
$$= \sum_z P(X_{n+m} = y | X_n = z, X_0 = x) P(X_n = z | X_0 = x)$$

donde de nuevo usamos propiedades de probabilidad condicional,

$$= \sum_z P(X_{n+m} = y | X_n = z) P(X_n = z | X_0 = x)$$

donde usamos la propiedad de Markov. Pero usando la fórmula (34), eso es igual a

$$= \sum_z Q^n(x,z) Q^m(z,y). \qquad \square$$

Urna de Ehrenfest Considere N bolillas distribuídas en dos urnas. Una bolilla es elegida al azar y es cambiada de urna. Este modelo representa el comportamiento de un gas que tiene N moléculas ocupando dos containers. Cual es la cadena de Markov que describe esta evolución temporal?

El espacio de estados es $S = \{0, 1, \ldots, N\}$ que describe el número de bolillas en la primera urna. Si en un momento hay k bolillas en la primera urna, las transiciones posibles son para $k-1$ (si $k > 0$) o para $k+1$ (si $k < N$) y las probabilidades de transición son

$$Q(k, k-1) = \frac{k}{N}, \qquad Q(k, k+1) = \frac{N-k}{N}$$

$$Q(x,y) = 0, \qquad \text{si } |x-y| > 1.$$

Si la primera urna tiene 4 bolillas y la segunda tiene 6, cual es la probabilidad que después de dos pasos haya 4 bolillas en la primera y 6 en la segunda?

$$Q^2(4,4) = Q(4,5)Q(5,4) + Q(4,3)Q(3,4) = \frac{6 \times 5 + 4 \times 7}{100}$$

Y cual es la probabilidad que despues de tres pasos haya 5 bolillas en cada urna? Hay que calcular $Q^3(4,5)$ que es igual a

$$Q(4,5)Q(5,6)Q(6,5) + Q(4,5)Q(5,4)Q(4,5) + Q(4,3)Q(3,4)Q(4,5)$$

13.4. Medidas invariantes

Considere una cadena de Markov en un espacio finito con matriz de transición Q para la cual existe un $k > 0$ tal que $Q^k(x,y) > 0$ para todo par de estados x, y. Suponga que existen los límites siguientes

$$\lim_n Q^n(x,y) =: \pi(y), \qquad \text{para todo par de estados } x, y$$

Si esto ocurre, decimos que la cadena olvida el valor inicial y que la distribución de X_n converge a π (convergencia en distribución) para cualquier estado inicial. Si escribimos

$$Q^{n+1}(x,y) = \sum_z Q^n(x,z) Q(z,y),$$

sacando límite en ambos miembros,

$$\pi(y) = \sum_z \pi(z) Q(z,y) \qquad \text{para todo } y$$

Estas son las **ecuaciones de balance**. Además, como $\sum_y Q^n(z,y) = 1$, para todo n, z y la suma es finita, tendremos que π es una probabilidad:

$$\sum_y \pi(y) = 1.$$

Es decir que la probabilidad π es un autovector a la izquierda de Q con autovalor 1: $\pi Q = \pi$. Una probabilidad π que satisface las ecuaciones de balance es llamada **medida invariante**. En particular,

$$\sum_x \pi(x) P(X_1 = y | X_0 = x) = \pi(y).$$

Es decir que si el estado inicial es aleatorio y con distribución π, entonces la distribución de la cadena en el instante 1 es también π. En general, para todo n, $\pi Q^n = \pi$:

$$\sum_x \pi(x) P(X_n = y | X_0 = x) = \pi(y)$$

O sea: si la distribución de X_0 es π, entonces la distribución de X_n es π para todo $n \geq 0$.

Ejemplo: lluvia. Las ecuaciones de balance son

$$\begin{aligned} \pi(0) &= 0{,}9\,\pi(0) + 0{,}2\,\pi(1), \\ \pi(1) &= 0{,}1\,\pi(0) + 0{,}8\,\pi(1) \\ \pi(0) + \pi(1) &= 1. \end{aligned}$$

Substituyendo las identidades $\pi(0) = \pi(0)(0{,}1 + 0{,}9)$ y $\pi(1) = \pi(1)(0{,}2 + 0{,}8)$ en los primeros términos, obtenemos que las ecuaciones de balance son equivalentes a

$$0{,}1\,\pi(0) = 0{,}2\,\pi(1); \qquad \pi(0) + \pi(1) = 1$$

cuya solución es

$$\pi(0) = \frac{0{,}2}{0{,}2 + 0{,}1} = \frac{2}{3}, \quad \pi(1) = \frac{0{,}1}{0{,}2 + 0{,}2} = \frac{1}{3}$$

Ejemplo: urna de Ehrenfest. Las ecuaciones de balance para $0 < k < N$ son:

$$\pi(k) = \pi(k+1)\frac{k+1}{N} + \pi(k-1)\frac{N-k+1}{N}, \quad 0 < k < N;$$

cuya solución es:

$$\pi(k) = \binom{N}{k} \left(\frac{1}{2}\right)^N$$

Teorema de existencia y unicidad de la medida invariante *Si el espacio de estados es finito y alguna potencia k de la matriz de transición tiene todas las entradas positivas, entonces la medida invariante existe y es única.*

Este teorema es un caso particular del teorema de Perron-Frobenius del Álgebra. Daremos una demostración probabilística basada en la ley de grandes números.

13.5. Ley de grandes números para cadenas de Markov

Medidas empíricas

Fijamos el estado inicial $X_0 = x$ y definimos

$$N_n(x,y) := \sum_{j=1}^{n} \mathbf{1}\{X_j = y\},$$

el número de visitas a y de la cadena empezando con $X_0 = x$ y la *distribución empirica*

$$\hat{Q}_n(x, y) := \frac{N_n(x, y)}{n}$$

Son variables aleatorias que indican la proporción de visitas al estado y hasta el instante n para la cadena que empieza en x. Para cada x y n fijos $\hat{Q}_n(x, \cdot)$ es una probabilidad aleatoria: $\sum_y \hat{Q}_n(x, y) = 1$. Calculemos la esperanza de $\hat{Q}$:

$$E\hat{Q}_n(x, y) = \frac{1}{n}\sum_{j=1}^{n} E(\mathbf{1}\{X_j = y\}|X_0 = x) = \frac{1}{n}\sum_{j=1}^{n} Q^j(x, y) \tag{36}$$

Defina el instante de la primera vuelta a t por

$$\tau(y) := \text{mín}\{n \geq 1 : X_n^y = y\}$$

donde X_n^y es la cadena con estado inicial $X_0^y = y$.

Teorema 22. *Sea X_n una cadena de Markov en un espacio de estados finito con matriz de transición Q. Asuma que existe $k > 0$ tal que $Q^k(x, y) > 0$ para todo par de estados x, y. Entonces para cada y,*

$$\lim_n \hat{Q}_n(x, y) = \frac{1}{E\tau(y)}, \qquad \textit{c.s.} \tag{37}$$

que no depende de x. Definiendo $\pi(y) := 1/E\tau(y)$, tenemos que π es la única medida invariante para la cadena.

Dem

Defina el instante de la j-ésima visita a y por $T_j(x, y) = \text{mín}\{k \geq 1 : N_k(x, y) = j\}$. Así,

$$N_n(x, y) = j \quad \text{ si y sólo si } \quad T_j(x, y) \leq n < T_{j+1}(x, y). \tag{38}$$

$$T_j(x, y) = T_1(x, y) + \tau_2(y) + \cdots + \tau_j(y) \tag{39}$$

donde $\tau_i(y)$ son iid con la misma distribución que $\tau(y) = T_1(y, y)$.

Como por hipótesis, a cada k pasos la cadena tiene por lo menos probabilidad $p := \text{mín}_{x,y} Q^k(x, y) > 0$ de visitar y,

$$P(T_1(x, y) > kn) \leq (1 - p)^n. \tag{40}$$

Por lo tanto, $\tau_i(y)$ tienen media y varianza finitas. Por la ley fuerte de grandes números tenemos

$$\lim_{j\to\infty} \frac{T_j(x, y)}{j} = E\tau(y), \qquad \text{c.s.} \tag{41}$$

Usando (38) cuando $N_n = j$, tenemos

$$\frac{T_j(x, y)}{j} \leq \frac{n}{N_n(x, y)} \leq \frac{T_{j+1}(x, y)}{j}. \tag{42}$$

Por otro lado $N_n(x, y) \to \infty$, caso contrario habría un $\tau_j(y) = \infty$, pero esto no puede ser por (40). Por lo tanto, sacando límites en (42), y recordando que $\pi(y) = 1/E\tau(y)$,

$$\pi(y) \leq \liminf_n \frac{N_n(x, y)}{n} \leq \limsup_n \frac{N_n(x, y)}{n} \leq \pi(y).$$

Esto demuestra (37).

Por el teorema de convergencia acotada, si el límite c.s. existe, entonces también existe el límite de las esperanzas:

$$\lim_n E\hat{Q}_n(x, y) = \pi(y) \tag{43}$$

y usando

$$E\hat{Q}_{n+1}(x,y) = \sum_z E\hat{Q}_n(x,z)Q(z,y) + O(1/n), \tag{44}$$

que demostraremos más abajo, tenemos que el límite π satisface las ecuaciones de balance.

Unicidad. Supongamos que exista otra medida invariante π' que satisface las ecuaciones de balance $\pi'Q = \pi'$. Entonces,

$$\pi'(y) = \sum_x \pi'(x)\frac{1}{n}\sum_{k=1}^{n} Q^k(x,y) \to_n \sum_x \pi'(x)\pi(y) = \pi(y),$$

usando (36) y (43). □

Demostracion de (44)

$$\begin{aligned} E\hat{Q}_{n+1}(x,y) &= \frac{1}{n+1}\sum_{j=1}^{n} Q^{j+1}(x,y) + \frac{1}{n+1}Q(x,y) \\ &= \frac{n}{n+1}\frac{1}{n}\sum_{j=1}^{n}\sum_z Q^j(x,z)Q(z,y) + \frac{1}{n+1}Q(x,y) \end{aligned}$$

De donde

$$E\hat{Q}_{n+1}(x,y) = \sum_z E\hat{Q}_n(x,z)Q(z,y) + O(1/n)$$

Uso de simulaciones para estimar π Una forma de estimar π es simular la cadena de Markov por un intervalo de tiempo de tamaño n "grande" y usar las distribuciones empíricas

$$\hat{Q}_n(x,y) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{X_k = y\}$$

como aproximación de $\pi(y)$. Hay una teoría que permite calcular los errores al hacer esta aproximación.

13.6. Aplicación. Algoritmo PageRank

Recomiendo leer el artículo de Wikipedia *PageRank*.

Consideramos un grafo orientado $\mathcal{G} = (V, E)$, donde $V =$ es un conjunto finito de vertices y $E \subset \{(x,y) : x, y \in V)$ es un conjunto de aristas orientadas. Vértices representan páginas web. Una arista orientada (x,y) indica que en la página x hay un link a la página y.

Queremos ranquear los vértices usando solamente la estructura del grafo. Una primera idea es usar el número de aristas que llegan a un vértice $y \in V$ y proponer el ranqueador

$$R_1(y) = \sum_{x\in V} a(x,y)$$

donde $a(x,y) = \mathbf{1}\{(x,y) \in E\}$ vale 1 cuando hay una arista que va de x a y. Pero esto le da mucho peso a los vértices que emanan muchas aristas. Para compensar, definimos el número de aristas que salen del vértice x por

$$a(x) = \sum_y a(x,y)$$

y dividiendo por este número obtenemos el segundo ranqueador:

$$R_2(y) = \sum_{x\in V} \frac{a(x,y)}{a(x)}$$

pero así todos los vértices que tienen el mismo número de aristas salientes envían el mismo peso, independientemente de las aristas entrantes. Mejor sería que cada vértice enviara un peso proporcional a su importancia, medida por las aristas que entran. Esto sugiere un tercer ranqueador:

$$R_3(y) = \sum_{x \in V} R_3(x) \frac{a(x,y)}{a(x)}$$

O sea que R_3 es la solución de un sistema de $|V|$ ecuaciones, una para cada vértice del grafo. Usando la notación

$$\pi = R_3, \qquad Q(x,y) = \frac{a(x,y)}{a(x)},$$

el tercer ranqueador satisface $\pi(y) = \sum_{x \in V} \pi(x)\, Q(x,y)$, las ecuaciones de balance para una cadena de Markov que se describe así:

"Cuando la cadena se encuentra en el vértice x, elige al azar, uniformemente, una de las flechas que salen de x y salta al extremo y de esa flecha"

Esta cadena no satisface la condición $Q^k > 0$ para algún k porque hay muchas páginas que no tienen links salientes. Por eso se propone una nueva matriz

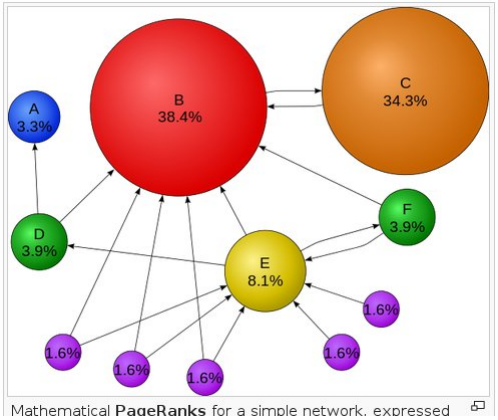
$$P := dQ + (1-d)\frac{1}{|V|}\mathbb{I}$$

donde $d \in [0,1]$ y $\mathbb{I}$ es una matriz de la misma dimensión de Q cuyas entradas son todas 1. Esta cadena opera así: "cuando está en el estado x, con probabilidad d elige una flecha de las salientes con la matriz Q, con probabilidad $(1-d)$ elige un vértice al azar en toda la web". La idea es que cada navegante al cabo de un tiempo deja de clicar links y se va o elige una página al azar en la red. El número de clicks que hace antes de saltar a un link aleatorio en la red es una variable aleatoria geométrica con parámetro $(1-d)$. Si $d = 0{,}85$, el número medio de clicks antes de aburrirse será $1/0{,}15 = 6{,}66$. El parámetro d se llama damping factor, o factor de amortiguamiento. El ranqueador será la solución de

$$R_4(y) = \sum_{x \in V} R_4(x)\Big(d\frac{a(x,y)}{a(x)} + (1-d)\frac{1}{|V|}\Big) \tag{45}$$

Vemos en http://news.netcraft.com/archives/category/web-server-survey/, que en octubre de 2015 hay más de 800 millones de páginas en la web. Con ese número de estados, el cálculo de la medida invariante en forma exacta es fisicamente imposible por el momento. La matriz está sea casi toda constituída de ceros porque cada página tiene links a unas pocas decenas o centenas de otras páginas. Cada fila de la matriz tiene tamaño 800 millones pero sólo unas pocas entradas son positivas.

Para estimar la medida invariante π para la matriz P se usa la ley de grandes números para cadenas de Markov, Teorema 22, que podemos hacer porque $P^1 \geq d\,\mathbb{I}$.



Medida invariante correspondiente a la matriz P con $d = 0{,}85$. Fuente PageRank de Wikipedia. Mathematical PageRanks for a simple network, expressed as percentages. (Google uses a logarithmic scale.)

Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85 % likelihood of choosing a random link from the page they are currently visiting, and a 15 % likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1 % of the time. (The 15 % likelihood of jumping to an arbitrary page corresponds to a damping factor of 85 %.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.

Se envía un robot que circula por los vértices de acuerdo a una cadena de Markov X_k con matriz de transición P por n pasos y se estima $\pi(y)$ con la medida empírica temporal

$$\hat{P}_n(x,y) := \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{X_k = y\}$$

o simplemente se usa $\hat{P}_n$ como ranqueador.

Otra posibilidad, que no discutiremos aquí, es estimar la medida invariante π para P por una fila cualquiera de P^n para n relativamente grande. Parece que $n = 50$ funciona bastante bien.

Paseos aleatorios

Contando caminos Un **camino** de longitud n es un vector $(s_0, s_1, \ldots, s_n)$,

$$s_k = x_1 + \cdots + x_k$$

donde los incrementos $x_i \in \{-1, 1\}$.

Hay 2^n caminos de longitud n. Si $s_0 = 0$ y $s_n = x$, entonces los a incrementos positivos y los b incrementos negativos deben satisfacer:

$$a + b = n, \qquad a - b = x.$$

Es decir:

$$a = \frac{n+x}{2}, \qquad b = \frac{n-x}{2}.$$

Así, $N_{n,x}$ el número de caminos de longitud n que van de 0 a x es

$$N_{n,x} = \binom{a+b}{a} = \binom{a+b}{b}$$

Consideraremos $N_{n,x} = 0$ cuando no se puede alcanzar x en n pasos.

Ejemplo Elecciones. Supongamos que en una elección el candidato A saca a votos y el candidato B saca b votos, con $a > b$ (es decir A gana la elección).

Cual es la probabilidad que durante todo el escrutinio A esté por delante de B?

Podemos representar la ventaja de A por un camino: cada vez que sale un voto para A sumamos 1 y cada vez que sale un voto para B restamos 1. O sea que $x_i = 1$ si el i-ésimo voto computado sale para A y $x_i = -1$ en caso que sea para B. La ventaja de A después de computar el k-ésimo voto es

$$s_k = x_1 + \cdots + x_k$$

A lidera todo el escrutinio si para todo $0 < k \leq n$,

$$s_1 > 0, s_2 > 0, \ldots, s_k > 0.$$

Asumimos que todos los posibles caminos de tamaño n que terminan en $a - b$ son igualmente probables. (todas las permutaciones de los votos son igualmente probables)

Principio de reflexión

Considere puntos espacio-temporales (k, x) y (n, y).

$0 \leq k < n$, $x > 0$, $y > 0$.

El punto **reflejado** de (k, x) es $(k, -x)$

Consideraremos caminos que van de (k, x) a (n, y).

Principio de reflexión *El número de caminos que van de (k, x) a (n, y) que toca o cruza el eje de las absisas es igual al número de caminos que van de $(k, -x)$ a (n, y).*

Dem Considere un camino $x = s_k, s_{k+1}, \ldots, s_n = y$ que toque el eje de las absisas. Sea T el primer instante en que eso sucede:

$$T = \text{mín}\{i \in [k, n] : s_i = 0\}$$

El camino

$$-x = -s_k, -s_{k+1}, \ldots, -s_{T-1}, 0, s_{T+1}, \ldots, s_n = y$$

va de $(k, -x)$ a (n, y).

Como las secciones $(k, x), \ldots, (t, 0)$ y $(k, -x), \ldots, (t, 0)$ son reflejadas una de la otra, existe una biyección entre esos dos pedazos. Esto implica que el número de caminos es el mismo. □

Lema (del escrutinio) ***Sean n y x enteros positivos. Hay exactamente $\frac{x}{n}N_{n,x}$ caminos $(s_1, \ldots, s_n = x)$ desde el origen a (n, x) tal que $s_1 > 0, \ldots, s_n > 0$.***

Dem Claramente hay tantos caminos admisibles como caminos desde $(1, 1)$ a (n, x) que no tocan el eje de las absisas. Por el lema de la reflexión, ese número es

$$N_{n-1,x-1} - N_{n-1,x+1} = \binom{a+b-1}{a-1} - \binom{a+b-1}{a}$$

con a y b satisfaciendo que $a+b = n$ y $a-b = x$. Una cuenta muestra que ese número es igual a $\frac{x}{n}N_{n,x}$. □

Paseos aleatorios son cadenas de Markov Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes con distribución

$$P(X_i = 1) = \frac{1}{2}, \qquad P(X_i = -1) = \frac{1}{2}.$$

Se define **paseo aleatorio** al proceso

$$S_n = X_1 + \cdots + X_n, \quad n \geq 0$$

S_n es una **cadena de Markov** con transiciones

$$q(x, x+1) = \frac{1}{2}, \quad q(x, x-1) = \frac{1}{2}.$$

Así, la probabilidad que el paseo esté en x en el instante n es

$$p_{n,x} = P(S_n = x) = \binom{n}{\frac{n+x}{2}} 2^{-n}$$

(se interpreta como 0 si $\frac{n+x}{2}$ no es un entero entre 0 y n.)

Una **vuelta al origen** ocurre en el instante $2k$ si $S_{2k} = 0$. La vuelta sólo puede ocurrir en instantes pares.

Definimos $u_{2k} = P(S_{2k} = 0)$.

$$u_{2k} = \binom{n}{\frac{k}{2}} 2^{-2k}$$

Ejercicio Use la aproximación de Stirling para probar que

$$u_{2k} \sim \frac{1}{\sqrt{\pi k}}$$

Eso quiere decir que

$$\lim_{k \to \infty} u_{2k}\sqrt{\pi k} = 1$$

El TCL nos dice que

$$\lim_{n} P(S_n \leq r\sqrt{n}) = \phi(r)$$

donde ϕ es la función de distribución acumulada de la Normal standard.

El **primer retorno al origen** ocurre en el instante $2k$ si

$$S_1 \neq 0, \ldots, S_{2k-1} \neq 0, S_{2k} = 0$$

y su probabilidad se denota f_{2k}.

Lema ***Las probabilidades*** u_{2k} ***y*** f_{2k} ***se relacionan por***

$$u_{2n} = f_2 u_{2n-2} + f_4 u_{2n-4} + \cdots + f_{2n} u_0$$

Dem Use el teorema de la probabilidad total. □

Sea $T := \min\{n > 0 : S_n = 0\}$ instante del primer retorno al origen.

Lema *Sea* $n > 0$*, entonces*

$$P(T > 2n) = P(S_{2n} = 0)$$

Dem Por simetría,

$$P(T > 2n) = P(S_1 > 0, \ldots, S_{2n} > 0) + P(S_1 < 0, \ldots, S_{2n} < 0)$$

$$= 2P(S_1 > 0, \ldots, S_{2n} > 0)$$

Por el teorema de la probabilidad total:

$$P(S_1 > 0, \ldots, S_{2n} > 0) = \sum_{x \geq 1} P(S_1 > 0, \ldots, S_{2n-1} > 0, S_{2n} = 2x)$$

Por el lema de reflexión,

$$P(S_1 > 0, \ldots, S_{2n-1} > 0, S_{2n} = 2x)$$

$$= 2^{-2n}(N_{2n-1,2x-1} - N_{2n-1,2x+1}) = \frac{1}{2}(p_{2n-1,2x-1} - p_{2n-1,2x+1})$$

Sumando (telescopicamente),

$$\sum_{x \geq 1} \frac{1}{2}(p_{2n-1,2x-1} - p_{2n-1,2x+1}) = \frac{1}{2} p_{2n-1,1} = \frac{1}{2} u_{2n} \qquad \square$$

Máximo El **máximo** M_n está definido por

$$M_n(S_0, \ldots, S_n) = \max\{S_0, \ldots, S_n\}$$

Lema *Sea* y *un entero tal que* $n \geq y > 0$*. La probabilidad de un camino de* $(0,0)$ *a* $(2n,0)$ *con un máximo mayor o igual a* y *es igual a* $p_{2n,2y} = P(S_{2n} = 2y)$*.*

Dem Queremos calcular $P(M_{2n} \geq y, S_{2n} = 0)$. El número de caminos de $(0,0)$ a $(2n,0)$ que tocan o cruzan y es igual al número de caminos de $(0,y)$ a $(2n,y)$ que tocan 0. Por el Lema de reflexión, ese número es igual a $N_{2n,2y}$. Multiplicando por 2^{-2n}, obtenemos

$$P(M_{2n} \geq y, S_{2n} = 0) = p_{2n,2y}. \qquad \square$$

Observe que

$$p_{2n,2y} = \binom{2n}{\frac{2n+2y}{2}} = \binom{2n}{n+y}$$

Lema

$$\lim_{n \to \infty} P\big(M_{2n} \geq b\sqrt{2n} \;\big|\; S_{2n} = 0\big) = e^{-2b^2}$$

Dem Dividiendo la expresión obtenida para $p_{2n,2y}$ por $p_{2n,0} = \binom{2n}{n} 2^{-2n}$, cancelan los $(2n)!$ y las potencias de 2 y obtenemos

$$P(M_{2n} \geq y | S_{2n} = 0) = \frac{p_{2n,2y}}{p_{2n,0}} = \frac{n!\, n!}{(n-y)!\,(n+y)!}$$

$$= \frac{n(n-1)\ldots(n-y+1)}{(n+y)(n+y-1)\ldots(n+1)}$$

dividiendo cada uno de los términos del denominador por el el correspondiente término del numerador, obtenemos

$$= \Big(\Big(1+\frac{y}{n}\Big)\Big(1+\frac{y}{n-1}\Big)\dots\Big(1+\frac{y}{n-y+1}\Big)\Big)^{-1}$$

Substituyendo $y = b\sqrt{2n}$, y

$$= \Big(\Big(1+\frac{b\sqrt{2}}{\sqrt{n}}\Big)\Big(1+\frac{b\sqrt{2}}{\sqrt{n}-\frac{1}{\sqrt{n}}}\Big)\dots\Big(1+\frac{b\sqrt{2}}{\sqrt{n}-\frac{b\sqrt{2}+1}{\sqrt{n}}}\Big)\Big)^{-1}$$

$$\sim \Big(1+\frac{b\sqrt{2}}{\sqrt{n}}\Big)^{-b\sqrt{2}\sqrt{n}} \to e^{-2b^2} \qquad \square$$

14. Inferencia estadística - Estimación puntual

Guerra de encuestas Noticias de los diarios sobre encuenstas pre-electorales:

Opinaia Es una de las pocas consultoras que mide de modo online y tiene en su haber dos grandes aciertos: uno, en la última elección bonaerense; el otro, en el comicio para jefe de Gobierno porteño en 2015. Ahora presentó un sondeo con 1.718 casos en Provincia, con un margen de error de ±2 %. Bullrich 39,5, Cristina 36,0.

UAI, Taquion y Axonier 850 encuestados bonaerenses. Bullrich 35,2, Cristina 32,2. Margen de error ±3,4 %. Es decir que se lo podría incluir en la categoría de "empate técnico".

Opina Argentina Otra de las consultoras que pronosticó con mucha exactitud los resultados de las PASO en la Provincia. Ahora, presentó un "informe de coyuntura política", en base a 1.200 casos, relevados entre el 3 y el 5 de este mes. Bullrich 38, Cristina 35.

Otra encuesta 3.000 casos, colectados el 5 y 6 de este mes, y con un margen de error de ±2 %. Bullrich 39, Cristina 35.

Resultados Abiertas las urnas, quedó este resultado: Bullrich 42,18, Cristina 36,25.

14.1. Estimación puntual

Esta Sección está basada en las notas de Ana Bianco y Elena Martinez . Para obtener una estimación de la proporción de p de votantes por un candidato antes de una elección se realiza una encuesta. La encuesta consiste en tomar una muestra de electores (aleatoria en el sentido que cada posible elector tiene la misma probabilidad de entrar en la muestra) y estimar p por la proporción muestral $\hat{p}$.

Ese procedimiento se basa en un **modelo**: se considera una variable aleatoria X Bernoulli con **parámetro** p y una *muestra aleatoria* $\underline{X} = (X_1, \dots, X_n)$, conformada por n variables aleatorias independientes con la misma distribución de X. En este caso $X_i = 1$ si el i-ésimo elector de la muestra vota por el candidato, caso contrario es 0.

La **proporción muestral** es la variable aleatoria

$$\hat{p}_n(\underline{X}) := \frac{X_1 + \dots + X_n}{n}$$

que será llamado *estimador* de p. Tiene sentido porque cuando n crece $\hat{p}_n(\underline{X}) \to p$ por la ley de grandes números.

Después de realizada la muestra los valores observados se denotan $\underline{x} = (x_1, \dots, x_n)$ y $\hat{p}(\underline{x})$ es una *estimativa* de p.

El **error** cometido al estimar p por $\hat{p}_n(\underline{X})$ es

$$|\hat{p}_n(\underline{X}) - p|$$

que también es aleatorio. Conociendo la distribución del error, podremos hacer afirmaciones sobre la bondad de nuestra estimación.

Parámetros Así como la Bernoulli depende del *parámetro* p, otras distribuciones de probabilidad dependen de cierto número de parámetros. Por ejemplo: Poisson depende de λ, Normal depende de μ y σ^2, Binomial depende de n y p, etc. Llamaremos Θ el espacio de parámetros y $\theta \in \Theta$ un parámetro, que puede ser un vector, como en el caso de la Normal $\theta = (\mu, \sigma^2)$.

Estimación puntual paramétrica

Sea $X = X_\theta, \theta \in \Theta$ una familia de variables aleatorias con distribución $X_\theta \sim F_\theta$.

Usaremos la notación

$$E_\theta g(\underline{X}) \tag{46}$$

para denotar la esperanza de $g(X_1, \ldots, X_n)$ cuando $\underline{X}$ es una muestra de X_θ la variable con distribución F_θ.

Un **estimador puntual** de θ es una función de la muestra de X que se denota

$$\hat{\theta} = \hat{\theta}(\underline{X})$$

Cuando el experimento es realizado, la muestra observada se denota con minúsculas $\underline{x}$ y $\hat{\theta}(\underline{x})$ se llama *estimativa*.

Ejemplo. En el caso de las encuestas electorales, consideremos p = probabilidad de votar en Bullrich; este es el parámetro desconocido que queremos estimar. Opinaia hizo una muestra de tamaño $n = 1718$ y $\hat{p} = 0{,}395$ es la proporción observada de intenciones de voto por Bullrich.

Métodos de estimación puntual

Vamos a estudiar algunas maneras de construir estimadores.

Método de momentos: Se buscan los valores de los parámetros que permiten igualar los momentos muestrales a los momentos poblacionales.

Sea $X = X_\theta$ una variable aleatoria con distribución F_θ, $\theta \in \Theta$.

Sea EX_θ^k el **momento** de orden k de X. Es una función de θ que llamamos g_k:

$$EX_\theta^k = g_k(\theta)$$

Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra de X.

Definimos el **momento muestral** de orden k por:

$$\frac{\sum_{i=1}^n X_i^k}{n}$$

Cuando la muestra observada es $(x_1, \ldots, x_n)$, los momentos observados de orden k son

$$\frac{\sum_{i=1}^n x_i^k}{n}$$

Suponga que $\theta = (\theta_1, \ldots, \theta_m)$. Es decir $\Theta = \mathbb{R}^m$.

Defina $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$ los parametros que se obtienen al igualar los m primeros momentos muestrales a los momentos poblacionales correspondientes. Más precisamente, $\hat{\theta}_1, \ldots, \hat{\theta}_m$ es la solución de las ecuaciones

$$g_k(\theta_1, \ldots, \theta_m) = \frac{\sum_{i=1}^n x_i^k}{n}, \quad k = 1, \ldots, m.$$

$(\theta_1, \ldots, \theta_m)$ son incógnitas y $(x_1, \ldots, x_n)$ son datos. Es decir que $\hat{\theta}_i = \hat{\theta}_i(x_1, \ldots, x_n) = \hat{\theta}_i(\underline{x})$ es una función de la muestra observada.

Substituyendo $\underline{x}$ por $\underline{X} = (X_1, \ldots, X_n)$, tenemos las variables aleatórias $\hat{\theta}_i(\underline{X})$ que se llaman **estimadores de momentos** de $(\theta_1, \ldots, \theta_m)$.

Ejemplo 1. $X \sim$ Exponencial(λ). Un parámetro, una ecuación:

$$EX = \bar{X}_n$$

Como $EX = 1/\lambda$, la ecuación queda

$$\frac{1}{\lambda} = \bar{X}_n.$$

De donde $\hat{\lambda} = 1/\bar{X}_n$.

Ejemplo 2. $X \sim \text{Gama}(\alpha, \lambda)$. Dos parametros, dos ecuaciones:

$$EX = \bar{X}_n, \qquad EX^2 = \frac{\sum_{i=1}^n X_i^2}{n}$$

Como $EX = \frac{\alpha}{\lambda}$ y $EX^2 = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2}$, las ecuaciones quedan

$$\frac{\alpha}{\lambda} = \bar{X}_n, \qquad \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2} = \frac{\sum_{i=1}^n X_i^2}{n}$$

De aqui se despejan λ y α:

$$\hat{\lambda} = \frac{\bar{X}}{\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2}, \qquad \hat{\alpha} = \frac{\bar{X}^2}{\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2}.$$

Ejemplo 3. $U \sim \text{Uniforme}[0, \theta]$. Un parametro, una ecuación:

$$EX = \bar{X}_n$$

como $EX = \frac{\theta}{2}$, la ecuación queda

$$\frac{\theta}{2} = \bar{X}_n.$$

Despejando θ obtenemos $\hat{\theta} = 2\bar{X}_n$.

Ejemplo 4. No siempre se puede usar el primer momento. Si X es Uniforme$[-\theta, \theta]$, $EX = 0$ no depende de θ, por eso hay que usar el segundo momento:

$$EX^2 = \frac{\sum_{i=1}^n X_i^2}{n}$$

como $EX^2 = \frac{4\theta^2}{12} = \frac{\theta^2}{3}$, la ecuación queda

$$\frac{\theta^2}{3} = \frac{\sum_{i=1}^n X_i^2}{n}.$$

Despejando θ, el estimador queda

$$\hat{\theta} = \sqrt{3\frac{\sum_{i=1}^n X_i^2}{n}}.$$

Método de máxima verosimilitud: Fisher en 1920.

Hallar los valores de los parámetros que maximizan la probabilidad de obtener la muestra observada.

Ejemplo: Encuesta de intención de voto de 20 personas. Queremos estimar la probabilidad p de votar por un determinado candidato. $X \sim \text{Bernoulli}(p)$. $(x_1, \ldots, x_n)$ son los valores observados. La probabilidad de haber observado $(x_1, \ldots, x_n)$ es

$$P\big((X_1, \ldots, X_n) = (x_1, \ldots, x_n)\big) = \prod_i p^{x_i}(1-p)^{1-x_i}$$

Cual es el valor de p que maximiza esa probabilidad?

$$\begin{aligned}
&\arg\max_p \Big[\prod_i p^{x_i}(1-p)^{1-x_i}\Big] \\
&\quad = \arg\max_p \Big[\log p \sum_i x_i + \log(1-p)\sum_i (1-x_i)\Big]
\end{aligned}$$

Buscamos el punto crítico derivando en p:

$$\frac{dg(p)}{dp} = \frac{1}{p}\sum_i x_i - \frac{1}{1-p}\sum_i (1-x_i) = 0$$

de donde

$$\hat{p} = \frac{\sum_i x_i}{n}$$

Calculando la derivada segunda vemos que maximiza.

Definición de estimador de máxima verosimilitud Sea $X = X_\theta$ una familia de variables aleatorias con rango R con probabilidad puntual $p_\theta(.)$ o densidad conjunta f_θ que depende de parámetros $\theta \in \Theta$, el espacio de parámetros.

La **función de verosimilitud** $L : \Theta \times R^n \to [0,1]$ está definida para $\theta \in \Theta$ y $\underline{x} = (x_1, \ldots, x_n) \in R^n$ por

$$L(\theta, \underline{x}) = \begin{cases} p_\theta(x_1) \ldots p_\theta(x_n) & \text{caso discreto} \\ f_\theta(x_1) \ldots f_\theta(x_n) & \text{caso continuo} \end{cases}$$

$L(\theta, (x_1, \ldots, x_n))$ es la probabilidad de observar $(x_1, \ldots, x_n)$ cuando el parámetro es θ. Para cada elemento $\underline{x} \in R^n$ definimos $\hat{\theta}(\underline{x})$ como el argumento que maximiza $L(\theta, \underline{x})$:

$$\hat{\theta}(\underline{x}) := \arg\max_\theta L(\theta, \underline{x});$$

usualmente hay apenas un argumento que maximiza L. Substituyendo $\underline{x}$ por las variables $\underline{X} = (X_1, \ldots, X_n)$ obtenemos el estimador

$$\hat{\theta}(X_1, \ldots, X_n)$$

que es llamado *estimador de máxima verosimilitud*. Usualmente se escribe $L(\theta)$ en lugar de $L(\theta, \underline{x})$, subentendiendo la dependencia de $\underline{x}$.

Ejemplos

1. $(X_1, \ldots, X_n) \sim$ Exponencial(λ)

$$L(\lambda) = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)}$$
$$\log L(\lambda) = n \log \lambda - \lambda(x_1 + \cdots + x_n)$$

Derivando en λ e igualando a cero

$$\frac{dL}{d\lambda} = \frac{n}{\lambda} + (x_1 + \cdots + x_n) = 0$$

De donde

$$\hat{\lambda} = \frac{1}{\bar{x}_n}.$$

Verifique que es un máximo con la segunda derivada.

2. $(X_1, \ldots, X_n) \sim$ Normal(μ, σ^2)

$$L(\mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big(-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\Big)$$

Maximizarla equivale a maximizar los logaritmos. El resultado es:

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}.$$

3. $(X_1, \ldots, X_n) \sim$ Uniforme$(0, \theta)$

$$L(\theta) = \frac{1}{\theta^n}\prod_i I_{x_i \in [0,\theta]} = 0 I_{\theta < \text{máx}_i\, x_i} + \frac{1}{\theta^n} I_{\theta \geq \text{máx}_i\, x_i}$$

De donde $\hat{\theta} = \text{máx}_i\, x_i$

Propiedades de los estimadores

Dada una muestra $(X_1, \ldots, X_n)$ de $X \sim F_\theta$ y un estimador puntual $\hat{\theta}(\underline{X})$, la diferencia

$$\hat{\theta}(\underline{X}) - \theta$$

es el **error** de estimación. Este error es una variable aleatoria dado que es función de la muestra. El **sesgo** de un estimador es su error medio:

$$b(\hat{\theta}) := E_\theta \hat{\theta}(\underline{X}) - \theta$$

Definición: Un estimador $\hat{\theta}(\underline{X})$ de θ es **insesgado** si tiene sesgo nulo. Un estimador $\hat{\theta}$ de θ es **asintóticamente insesgado** si

$$\lim_n E_\theta \hat{\theta} = \theta$$

Ejemplos.

1. $X \sim Bernoulli(p)$. La proporción muestral $\hat{p}$ es un estimador insesgado de p:

$$E_p \hat{p} = p.$$

2. $X \sim Normal(\mu, \sigma)$. Es claro que $\hat{\mu} = \bar{X}$ es insesgado pero

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2$$

no es insesgado para σ^2 porque haciendo cuentas se obtiene

$$E_{\mu,\sigma^2} \hat{\sigma}^2 = \frac{n-1}{n} \sigma^2,$$

o sea que $\hat{\sigma}^2$ es estimador asintóticamente insesgado de σ^2. Las mismas cuentas dicen que S^2 es estimador insesgado de σ^2.

3. $X \sim Uniforme[0, \theta]$. El estimador de momentos de θ es $2\bar{X}$. Es insesgado: $E_\theta \bar{X} = \theta$. El EMV de θ es $M = \text{máx}_i\, X_i$. No es insesgado:

$$\begin{aligned} E_\theta M = \int_0^\theta P_\theta(M > x) dx &= \int_0^\theta (1 - P_\theta(M \leq x)) dx \\ = \int_0^\theta \Big(1 - \Big(\frac{x}{\theta}\Big)^n\Big) dx &= \theta - \frac{\theta^{n+1}}{(n+1)\theta^n} = \frac{n}{n+1}\theta \end{aligned}$$

pero es asintoticamente insesgado.

Estimadores de mínima varianza

Hay muchos estimadores insesgados de los parametros.

Al estimar μ de la normal, por ejemplo.

$\hat{\mu}_1 = \bar{X}$, $\hat{\mu}_2 = \frac{X_1 + X_2}{2}$, $\hat{\mu}_3 = X_4$ son todos estimadores insesgados.

Sus varianzas son

$$V\bar{X} = \frac{\sigma^2}{n}, \qquad V\Big(\frac{X_1 + X_2}{2}\Big) = \frac{\sigma^2}{2}, \qquad V\Big(X_4\Big) = \sigma^2.$$

parece natural elegir el estimador más preciso, es decir el de menor varianza.

Principio de estimación insesgada de mínima varianza: Entre todos los estimadores insesgados de θ , elegir el de menor varianza. El estimador resultante se denomina IMVU (insesgado de mínima varianza uniformemente). Existe una metodología que permite hallar estimadores IMVU en muchas situaciones.

Teorema *Sea X una variable con distribución $N(\mu, \sigma^2)$. Entonces $\bar{X}$ es estimador IMVU de μ.*

Así, si se tiene evidencia de que la muestra viene de una distribución Normal, parece conveniente usar $\bar{X}$ como estimador de μ.

Si los datos no son Normales este estimador podría llegar a ser una pésima elección.

Ejemplos

a. Normal (μ, σ^2)

b. Cauchy $f(x) = \frac{1}{\pi(1+(x-\mu)^2)}$

c. Uniforme en $[\mu - 1, \mu + 1]$

Consideremos los siguientes estimadores:

$$\hat{\mu}_1 = \bar{X}, \qquad \hat{\mu}_2 = \widetilde{X}, \qquad \hat{\mu}_3 = \frac{max\ X_i + \text{mín}\ X_i}{2}$$

a. Para la normal el IMVU es $\bar{X}$ por lo que elegimos $\hat{\mu}_1$.

b. Cauchy muy dispersa, mejor elegir $\widetilde{X}$ (mediana muestral)

c. Elegir $\hat{\mu}_3$ porque la distribución no tiene colas.

El **Error standard de un estimador** $\hat{\theta}$ es su desviación standard, es decir

$$\sigma(\hat{\theta}) = \sqrt{V_\theta(\hat{\theta})}$$

Si el error standard depende de parámetros desconocidos, éstos se reemplazan por un estimador y se obtiene el error standard estimado.

Ejemplo: $X \sim$ Normal (μ, σ^2). El EMV de μ es $\bar{X}$ y el error standard es

$$\sqrt{V_\mu \bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Como depende de σ, podemos estimarlo substituyendo σ por S^2.

$$\hat{\sigma}(\bar{X}) = \frac{S}{\sqrt{n}}$$

Error cuadrático medio (ECM) $\hat{\theta}$ estimador de θ.

Def:

$$\text{E}CM_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2$$

Tenemos la siguiente fórmula:

$$\begin{aligned} \text{E}CM_\theta(\hat{\theta}) &= E_\theta(\hat{\theta} - \theta)^2 \\ &= E_\theta(\hat{\theta} - E_\theta\hat{\theta} + E_\theta\hat{\theta} - \theta)^2 \\ &= V_\theta(\hat{\theta}) + (b_\theta(\hat{\theta}))^2 \end{aligned} \tag{47}$$

donde $b_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$ es el sesgo. Por lo tanto, si $\hat{\theta}$ es insesgado, el ECM coincide con la varianza de $\hat{\theta}$.

Principio de estimación de menor error cuadrático medio: Dados dos o más estimadores del parámetro, elegir el de menor ECM. Cuando son insesgados, esto corresponde a elegir al estimador de mínima varianza.

Este principio permite además seleccionar, entre un estimador insesgado y otro que no lo es, en base a la varianza y al sesgo. Si el estimador sesgado tiene una varianza mucho menor que el insesgado, podría ser preferible su uso.

Consistencia

Diremos que $\hat{\theta}_n(\underline{X})$ es un estimador **consistente** de θ si

$$\hat{\theta}_n(\underline{X}) \longrightarrow \theta, \quad \text{en probabilidad.}$$

Ejemplo Si X tiene media μ y varianza σ^2, entonces $\bar{X}_n$ es un estimador consistente de μ, por la ley de grandes números. Verifique que el estimador $(X_1 + X_n)/2$ no es consistente.

Lema 23. *Si un estimador es asintóticamente insesgado y su varianza va a cero, entonces es consistente.*

Dem: Por Chevichev tenemos

$$\begin{aligned} P_\theta(|\hat{\theta} - \theta| > \varepsilon) &\leq \frac{E_\theta(\hat{\theta} - \theta)^2}{\varepsilon^2} \\ &= \frac{1}{\varepsilon^2}\big(V_\theta(\hat{\theta}) + (b_\theta(\hat{\theta}))^2\big) \qquad \text{(por (47))} \\ &\underset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

porque $V_\theta(\hat{\theta}) \to 0$ por hipótesis y $b_\theta(\hat{\theta}) \to 0$ porque $\hat{\theta}$ es asintoticamente insesgado. □

Ejemplo $X \sim$ Uniforme$[0, \theta]$. Vimos que $\hat{\theta} = \max X_i$ es asintoticamente insesgado, porque calculamos $E_\theta(\hat{\theta}) = \frac{n}{n+1}\theta \to \theta$. La varianza del máximo de n uniformes es

$$V(\max_i X_i) = \frac{n}{(n+1)(n+2)^2}\theta^2 \to_n 0$$

Por lo tanto $\hat{\theta} = \max X_i$ es consistente.

Lema 24. *S^2 es un estimador consistente de la varianza poblacional.*

Dem

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum (X_i - \bar{X})^2 = \frac{1}{n-1}\sum (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \frac{1}{n-1}\Big(\sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\Big) = \frac{n}{n-1}\Big(\sum \frac{X_i^2}{n} - \bar{X}^2\Big). \end{aligned}$$

Como $\bar{X}_n \to \mu$, $(\bar{X}_n)^2 \to \mu^2$. Además, por la ley de grandes números,

$$\sum_i \frac{X_i^2}{n} \to E_{\mu,\sigma^2} X^2 = \mu^2 + \sigma^2.$$

Como $n/(n-1) \to 1$,

$$S_n^2 \to \mu^2 + \sigma^2 - \mu^2 = \sigma^2. \qquad \square$$

14.2. Intervalos de confianza

Vimos estimación puntual de un parámetro, y controlamos en algunos casos el error entre el estimador y el parámetro. Otro modo es reemplazar la estimación puntual por un intervalo de valores posibles para el parámetro.

Ejemplo Si $X \sim$ Normal(μ, σ^2) con μ desconocida y σ^2 conocida. Sabemos que $\bar{X}_n \sim$ Normal$(\mu, \sigma^2/n)$ y que

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{ Normal}(0, 1),$$

de donde, usando la tabla Normal, obtenemos

$$P\Big(-1,96 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1,96\Big) = 0,95$$

que equivale a

$$P(\bar{X} - 1,96\,\sigma/\sqrt{n} \leq \mu \leq \bar{X} - 1,96\,\sigma/\sqrt{n}) = 0,95 \tag{48}$$

Es decir que la probabilidad que el intervalo aleatorio

$$I := [\bar{X} - 1,96\,\sigma/\sqrt{n}, \bar{X} - 1,96\,\sigma/\sqrt{n}]$$

contenga μ es 0,95. Se trata de un intervalo de radio $1,96\,\sigma/\sqrt{n}$ y centro aleatorio $\bar{X}$. El intervalo I es el **intervalo de confianza para μ de confianza 0,95**.

La expresión (48) se puede interpretar así: "en 95 % de las muestras de tamaño n, el intervalo de radio $1{,}96\,\sigma/\sqrt{n}$ centrado en la media muestral $\bar{x}$ contiene a la media poblacional μ".

En general, la relación entre el radio del intervalo ε, la confianza γ y el tamaño de la muestra n está dada por la siguiente fórmula:

$$\varepsilon = z_\gamma\,\sigma/\sqrt{n} \tag{49}$$

donde z_γ está determinada por la fórmula $\Phi(z_\gamma) - \Phi(-z_\gamma) = \gamma$; ϕ es la acumulada de la normal standard.

Definición Sea X una variable aleatoria cuya distribución depende de un parámetro $\theta \in \mathbb{R}$ y $\underline{X} = (X_1, \ldots, X_n)$ una muestra de X. Si a y b son dos funciones de la muestra tales que

$$P(a(\underline{X}) \leq \theta \leq b(\underline{X})) = 1 - \alpha, \tag{50}$$

llamamos al intervalo $[a(\underline{X}), b(\underline{X})]$ el *intervalo de confianza a nivel* $1 - \alpha$ para θ.

Observaciones: 1) El intervalo $[a, b]$ es aleatorio ya que sus extremos son funciones de la muestra. La expresión (50) debe leerse "La probabilidad de que el intervalo (a,b) contenga al parámetro θ es $1 - \alpha$".

2) Una vez observada la muestra, el intervalo es también "observado" y ya no tiene sentido hablar de probabilidad, sino de "confianza" de que el intervalo contenga a θ. Como $(1 - \alpha)100\,\%$ de las muestras producirán intervalos que contienen a θ, esa es nuestra confianza de que el intervalo observado sea uno de esos.

Intervalos de confianza asintótico para p de la Bernoulli. Sea X Bernoulli(p) y $\hat{p}_n$ el estimador puntual de p. Queremos establecer la relación entre el tamaño de la muestra n, el radio del intervalo dado por el error ε y la confianza $1 - \alpha$, basados en la expresión

$$P(\hat{p}_n - \varepsilon < p < \hat{p}_n + \varepsilon) = 1 - \alpha$$

que equivale a

$$P(|\hat{p}_n - p| < \varepsilon) = 1 - \alpha$$

Standarizando obtenemos la expresión equivalente

$$P\Big(\frac{|\hat{p}_n - p|}{\sqrt{p(1-p)}/\sqrt{n}} < \frac{\varepsilon}{\sqrt{p(1-p)}/\sqrt{n}}\Big) = 1 - \alpha$$

Por el teorema del límite central, aproximadamente

$$P\Big(|Z| < \frac{\varepsilon}{\sqrt{p(1-p)}/\sqrt{n}}\Big) \approx 1 - \alpha$$

para $Z \sim \text{Normal}(0, 1)$. Aceptando la aproximación como identidad, obtenemos la siguiente relación:

$$z = \frac{\varepsilon}{\sqrt{p(1-p)}/\sqrt{n}} \tag{51}$$

donde $z = z_{(1-\alpha)/2}$ satisface $P(|Z| < z) = 1 - \alpha$. Para usar la tabla, observe que $P(|Z| < z) = 1 - \alpha$ es equivalente a $\phi(z) = 1 - \alpha/2$, con ϕ la acumulada de la Normal$(0, 1)$.

Preguntas:

1) Dado el error ε y el tamaño n de la muestra, cual es la confianza $1 - \alpha$ del intervalo obtenido?

2) Dado el error ε y la confianza $1-\alpha$ que deseamos, cual es el tamaño n que debe tener la muestra?

3) Dada la confianza $1-\alpha$ que deseamos que tenga el intervalo obtenido y el tamaño n de la muestra, cual es el error obtenido?

Respuestas: Use (51) y la desigualdad $p(1-p) \leq \frac{1}{4}$ para obtener:

1) Primero se obtiene z con la fórmula

$$z = \frac{\varepsilon}{\sqrt{p(1-p)}/\sqrt{n}} \geq 2\varepsilon\sqrt{n}$$

y de ahí $1-\alpha$ usando la tabla: $P(Z < z) = (1-\alpha/2)$. El intervalo obtenido con este z va a tener confianza $(1-\alpha)$, por lo menos.

2) Tenemos $1-\alpha$ y ε y despejamos n:

$$n = \frac{z^2 p(1-p)}{\varepsilon^2} \leq \frac{z^2}{2\varepsilon^2}$$

Obtenga z usando la tabla y use el menor n mayor o igual a $\frac{z^2}{2\varepsilon^2}$ para garantizar un intervalo de radio ε con confianza $1-\alpha$.

3) Ahora conocemos $1-\alpha$ y n y buscamos ε. Despeje en (51):

$$\varepsilon = \frac{z\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{z}{2\sqrt{n}}$$

tomando el peor caso. Obtenemos z a partir de $1-\alpha$ como antes y el intervalo de radio $\varepsilon = \frac{z}{2\sqrt{n}}$ va a tener confianza por lo menos $1-\alpha$.

Ejemplos. Encuestas Provincia de Buenos Aires 2017

Construya los intervalos de confianza a partir de los datos informados por las encuestadoras y calcule los coeficientes de confianza respectivos.

El intervalo de confianza $(1-\alpha)$ para la intención de voto p para cada candidato es

$$[\hat{p} - \varepsilon, \hat{p} + \varepsilon],$$

donde

$$\varepsilon = z_{1-\alpha}\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq z_{1-\alpha}\frac{1}{2\sqrt{800}}. = 0{,}03464$$

y $z_{1-\alpha}$ satisface

$$P(-z_{1-\alpha} < Z < z_{1-\alpha}) = 1-\alpha.$$

Intervalo de confianza asintótico para la media de variables con varianza conocida

Sea X una variable aleatoria con media μ (desconocida) y varianza σ^2 conocida.

Usamos que la distribución asintótica de

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}$$

es aproximadamente Normal(0,1) para obtener el siguiente intervalo de confianza asintótica $1-\alpha$:

$$\left[\bar{X} - z\frac{\sigma}{\sqrt{n}},\ \bar{X} + z\frac{\sigma}{\sqrt{n}}\right]$$

donde $P(Z < z) = 1-\alpha/2$

Distribución chi cuadrado $X \sim N(\mu, \sigma^2)$

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma} \sim \chi^2_n$$

La suma de n normales standard al cuadrado tiene distribución Chi cuadrado con n grados de libertad.

Además χ^2_n es la misma distribución que Gama$(\frac{n}{2}, \frac{1}{2})$.

La tabla está organizada así. Si $X \sim \chi^2_n$, entonces

$$P(X > \chi^2_{n,\alpha}) = \alpha. \tag{52}$$

Hay una linea para cada n de 1 a 30 y una columna para diversos valores de α que van dedecreciendo de 0,995 a 0,005.

Por ejemplo, si $X \sim \chi^2_{11}$,

$$P(X > 21{,}920) = 0{,}025. \tag{53}$$

Distribución t

Sean $Z \sim$ Normal$(0,1)$ y $U \sim$ Gama$(\frac{n}{2}, \frac{1}{2}) = \chi^2_n$;

Z y U variables independientes. Definimos

$$T = \frac{Z}{\sqrt{U/n}} \sim t_n$$

t de student con n grados de libertad. Tabulada. Campana como la normal pero colas pesadas.

La tabla está organizada así: si $X \sim t_n$, entonces

$$P(X > t_{n,\alpha}) = \alpha \tag{54}$$

Hay una linea para cada n de 1 a 100 y una columna para diversos valores de α que van dedecreciendo de 0,25 a 0,005. Por ejemplo, si $X \sim t_7$, entonces

$$P(X > 2{,}3646) = 0{,}025. \tag{55}$$

Proposición 25. *$X_1, \ldots, X_n$ muestra de Normal(μ, σ^2). Entonces:*

a) $\bar{X} \sim N(\mu, \sigma^2/n) \iff \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$

b) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

c) $\bar{X}$ *y* S^2 *son independientes*

d) $\sqrt{n}\frac{\bar{X}-\mu}{S} \sim t_{n-1}$

IC para la media de la Normal(μ, σ^2) con varianza desconocida:

En este caso,

$$P\Big(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\Big) = 1 - \alpha \tag{56}$$

donde $P(T_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$. Así, el IC de confianza $(1-\alpha)$ está dado por

$$\bar{X} \pm t_{n-1,\alpha/2} S/\sqrt{n}$$

IC para la varianza de la normal con media conocida:

Como

$$P\Big(\chi^2_{n,1-\alpha/2} \le \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma} \le \chi^2_{n,\alpha/2}\Big) = 1 - \alpha$$

Despejando, construímos el intervalo de confianza $1-\alpha$ para σ:

$$\Big[\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{n,1-\alpha/2}}\Big]$$

IC para la varianza de la normal con media desconocida:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Intervalo:

$$\Big[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\Big]$$

Método del pivote para obtener intervalos de confianza:

Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra de X cuya distribución depende de un parámetro θ. Supongamos que existe una función $T(\underline{X}, \theta)$ (es decir, una función de la muestra y del parámetro) cuya distribución no depende de θ ni de ningún otro parámetro desconocido. Entonces, para cada $\alpha > 0$ existen dos valores a y b tales que

$$P(a < T(\underline{X}, \theta) < b) = 1 - \alpha.$$

A partir de esta expresión es posible obtener un intervalo de confianza para θ. T es llamado **pivote**.

Ejemplo X Exponencial(λ).

$$X_1 + \cdots + X_n \sim \text{ Gama}(n, \lambda)$$

Como $\lambda X \sim$ Exponencial(1), se puede demostrar que

$$T = 2\lambda(X_1 + \cdots + X_n) \sim \chi^2_{2n}$$

Con eso se puede obtener un intervalo de confianza para λ con la tabla de la χ^2 con $2n$ grados de libertad. De la tabla de la Chi cuadrado:

$$P(\chi^2_{2n,1-\frac{\alpha}{2}} < T < \chi^2_{2n,\frac{\alpha}{2}}) = 1 - \alpha$$

se despeja λ para obtener el intervalo

$$P\Big(\frac{\chi^2_{2n,1-\frac{\alpha}{2}}}{2\sum X_i} < \lambda < \frac{\chi^2_{2n,\frac{\alpha}{2}}}{2\sum X_i}\Big) = 1 - \alpha$$

Ejemplo. Sea X Uniforme$[0, \theta]$. El EMV de θ es $\hat{\theta} = \max X_i$. La distribución de $T = \hat{\theta}/\theta$ no depende de θ. De hecho, la distribucion de X_i/θ es Uniforme$[0, 1]$ y la distribución de T es la del máximo entre n Uniforme$[0, 1]$. La acumulada de T es

$$F_T(x) = x^n, \quad x \in [0, 1]$$

Derivando obtenemos la densidad

$$f_T(x) = nx^{n-1} I_{[0,1]}, \quad x \in [0, 1]$$

Usando T como pivote, tenemos

$$P(a < T < b) = 1 - \alpha \tag{57}$$

obtenemos el siguiente intervalo de confianza $1 - \alpha$:

$$\Big[\frac{\max X_i}{b}, \frac{\max X_i}{a}\Big]$$

Cómo elegir a y b? Son soluciones de (57) para α fijo:

$$\int_a^b nx^{n-1} \mathbf{1}\{x \in [0, 1]\} dx = 1 - \alpha,$$

que tiene infinitas soluciones. Podemos buscar a y b que minimicen la longitud promedio del intervalo de confianza que es $(E\,\text{máx}\,X_i)[\frac{1}{a}-\frac{1}{b}]$. Es decir, a y b que satisfagan $0 \le a \le b \le 1$ y

$$\begin{aligned} b^n - a^n &= 1-\alpha \\ \frac{1}{a}-\frac{1}{b} &= \text{mín} \end{aligned} \tag{58}$$

Substituyendo la primera linea en la segunda, ese problema es equivalente a minimizar la siguiente función de b:

$$\frac{1}{\left(b^n-1+\alpha\right)^{1/n}} - \frac{1}{b} = \text{mín} \tag{59}$$

Esta es una función decreciente en b, que alcanza su minimo en $b=1$, que es el máximo valor posible de b. Substituyendo $b=1$ en (58), obtenemos $a=\alpha^{1/n}$.

15. Test de Hipotesis

En una isla hay dos tribus. Una de aborígenes amigables de altura media 170 cm y otra de caníbales de altura media 150 cm.

Al llegar a la isla un explorador encuentra 9 aborígenes y tiene que decidir si son amigables o caníbales.

La altura de los aborígenes encontrados es una variable aleatoria X (cm).

Asumimos $X \sim N(\mu, 100)$. Varianza conocida. μ desconocida, o es 150 o 170.

Necesitamos decidir entre una de las dos hipótesis:

$H_0 : \mu = 150$, los aborígenes son caníbales.

$H_1 : \mu = 170$, los aborígenes son amigables.

Obtenemos una muestra aleatoria $(X_1, \ldots, X_9)$ de X y calculamos su media muestral $\bar{X}_9$.

Regla de decisión: Decidimos que si $\bar{x}_9 > 160$, se rechaza H_0 y se desembarca en la isla. Por el momento 160 es un valor arbitrario.

En caso contrario, se acepta H_0 y se escapa lo más rápido posible.

Es decir testeamos la hipótesis H_0 con el criterio "si la media muestral está arriba de 160, la rechazamos; si no, la aceptamos".

Región crítica (o de rechazo) para $\bar{x}$ es el intervalo $(160, \infty)$.

Por ejemplo, si observamos $\bar{x} = 162$. Qué hacemos?

Como el valor observado está en la región crítica (es mayor que 160), rechazamos H_0.

Podemos cometer dos errores:

Error de tipo 1: Rechazar H_0 cuando H_0 es verdadera.

Error de tipo 2: Aceptar H_0 cuando H_0 es falsa.

Cual es la probabilidad de cometer el error de tipo 1?

Usaremos que bajo H_0 conocemos la distribución de $\bar{X}_9$.

La media muestral tiene distribución normal: $\bar{X}_9 \sim N(\mu, 100/9)$.

Cálculo de la probabilidad del error 1

$$\begin{aligned} \alpha = P(\text{error tipo 1}) &= P(\bar{X}_9 > 160 | H_0 \text{ verdadera}) \\ &= P(\bar{X} \ge 160 | \mu = 150) \end{aligned}$$

$$= P\Big(\frac{\bar{X}-150}{10/3} \geq \frac{160-150}{10/3}\Big|\mu = 150\Big)$$

pero, como bajo $\mu = 150$, $Z = (\bar{X}-150)/(10/3) \sim N(0,1)$,

$$= P(Z > 3) = 0{,}0013 \quad \text{(por la tabla)}$$

α es el **nivel de significancia** del test.

Entonces, si observamos 162, rechazamos y si observamos 157, no rechazamos H_0.

Qué quiere decir $\alpha = 0{,}0013$?

Informalmente: Que de cada 10000 muestras que provienen de una población con H_0 verdadera (es decir $\mu = 150$), rechazaremos (equivocadamente) H_0 en 13 de los tests.

Si van 10000 expediciones a esa isla y observan una muestra de 9 caníbales, habrá 13 que desembarcarán (y serán comidos por los caníbales).

Definición Dadas dos hipótesis H_0 y H_1 relativas a parámetros de la distribución de una variable aleatoria X, un **test** es una **regla de decisión** basada en un estadístico o función de una muestra de X y en una zona de rechazo, es decir un conjunto de valores para los cuáles se rechaza la hipótesis nula H_0.

En el ejemplo anterior el estadístico era $\bar{X}$ y la zona de rechazo el intervalo $[160, \infty)$.

La zona de rechazo es fija. Se puede fijar en función del error α que estamos dispuestos a cometer.

La regla de decisión es aleatoria, porque depende del valor del estadístico.

Podemos equivocarnos. Por ejemplo podemos rechazar H_0 aún siendo $\mu = 150$.

Es imposible construir tests en los cuáles estemos absolutamente seguros de tomar la decisión correcta

Tipos de error:

Tipo 1: Se rechaza Ho cuando Ho es cierta

Tipo 2: No se rechaza Ho cuando Ho no es cierta

$\alpha = P($ error tipo 1) Nivel de significancia.

$\beta = P($ error tipo 2)

¿Cómo se elige la zona de rechazo?

Elegiremos la zona de rechazo del test de manera que la probabilidad de error tipo 1 sea un valor α predeterminado.

En el ejemplo, para $\alpha = 0,05$, buscamos z tal que $\phi(z) = 1 - 0{,}05$ y rechazamos Ho si $\frac{\bar{X}-150}{10/3} > z$ que corresponde a $z = 1{,}64$ y

$$\bar{x} \geq 150 + 1{,}64\frac{10}{3} = 150 + 5{,}4 = 154,4$$

Para $\alpha = 0,05$ rechazamos si $\bar{x} \geq 154,4$.

***P*-valor** Otra manera de hacer el test es considerar un estadístico llamado P-valor.

Si estamos considerando el estadístico T y observamos $t_{\text{observado}}$, el P-valor es el α correspondiente a la región crítica para T cuyo extremo es $t_{\text{observado}}$.

En particular, para el ejemplo anterior con el estadístico $T = \bar{X}$, si se la muestra observada es $x_1, \ldots, x_n$ y la media muestral observada es $\bar{x} = \bar{x}_{\text{observado}} = 156$, el P-valor es

$$P\text{-Valor}(x_1, \ldots, x_n) = P(\bar{X} > \bar{x} \,|\, H_0)$$

$$= P(\bar{X}_9 > 156 \,|\, \mu = 150) = P(Z > 1{,}8) = 0{,}0359.$$

(por la tabla) Esto quiere decir que si hacemos un test con $\alpha < 0{,}0359$, no podremos rechazar H_0.

Substituyendo $(x_1, \ldots, x_n)$ por $(X_1, \ldots, X_n)$, obtenemos el estadístico $P(X_1, \ldots, X_n)$. El P-valor es una función de la muestra, por lo tanto es un estadístico.

Para rechazar H_0, el P-valor observado tiene que ser menor que el α deseado. O sea, la región crítica para el P-valor es $[0, \alpha]$.

Error tipo 2

Supongamos que en nuestro ejemplo, observamos una altura media 154 en la muestra de tamaño 9 y trabajamos con el test de nivel 0.05.

En este caso,

$$\bar{x} = 154 \leq 154{,}4$$

que está fuera de la región crítica $[154{,}4, \infty)$.

Por lo tanto **no rechazamos** H_0.

Podríamos estar cometiendo un error de tipo 2.

Por ejemplo, si los aborigenes observados no son caníbales y tienen altura media 160, ¿cuál es la probabilidad de cometer un error tipo II?

$$\begin{aligned} P(\text{error tipo 2}) &= P(\text{aceptar } H_0 \,|\, H_1 \text{ verdadera, con } \mu = 160) \\ &= P(\bar{X}_9 < 154{,}4 \,|\, H_1 \text{ verdadera, con } \mu = 160) \\ &= P\Big(\frac{\bar{X} - 160}{10/3} < \frac{154{,}4 - 160}{10/3} \Big| \mu = 160\Big) \\ &= P(Z < -1{,}68) = 1 - 0{,}9535 = 0{,}0465 \end{aligned}$$

(usando la tabla).

El error de tipo 2 es una **función** del valor alternativo de H_1 y de la región crítica.

En este caso $\beta(9{,}3) = 0{,}0465$. Depende de la región crítica y del valor alternativo de θ bajo H_1.

Analogía con el sistema de justicia

Suponga que alguien es acusado de un crimen. La hipótesis nula es que la persona es inocente. La hipótesis alternativa es que el acusado es culpable. El test de hipótesis es un juicio con pruebas presentadas por las dos partes. Una vez consideradas las presentaciones de la acusación y la defensa, el jurado toma la decisión de "culpable" o "no culpable". El juicio nunca declara *inocente* al acusado, a lo sumo concluye que las pruebas presentadas no son suficientes para declararlo culpable. El objetivo del juicio es determinar si hay pruebas suficientes para declararlo culpable.

El error de tipo 1 corresponde a declarar culpable a un inocente. El error de tipo 2 es liberar a una persona culpable. El error de tipo 1 es el más serio ("somos todos inocentes hasta que se demuestre lo contrario"). Por ese motivo se busca que la probabilidad de ese error sea muy chica. En juicios criminales, lo usual es declarar culpable al acusado cuando hay poco espacio para la duda.

Función de potencia de un test, Fijada la región crítica, se llama

potencia $\pi(\mu)$ a la función que da la **probabilidad de rechazar la hipótesis nula cuando el valor verdadero del parámetro es** μ.

Utilizando la función de potencia es posible obtener una expresión general para los dos tipos de errores, pues

$$\pi(\mu) = \alpha(\mu) I\{\mu \in H_0\} + (1 - \beta(\mu)) I\{\mu \in H_1\}$$

Tipos de hipotesis

Las hipótesis alternativas pueden ser unilaterales o bilaterales. Las regiones de rechazo dependen del tipo de test.

Ejemplo, el test para μ de la normal con σ^2 conocida.

Hay tres posibles tests para μ:

1) H_0: $\mu = \mu_0$, H_1: $\mu < \mu_0$; (contra menor)

2) H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$; (contra mayor)

3) H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$; (bilateral)

Usamos el estadístico

$$T = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma},$$

Como bajo H_0, $T \sim N(0,1)$, las regiones de rechazo a nivel α son, respectivamente:

1) $RC = (-\infty, -z_\alpha]$

2) $RC = [z_\alpha, \infty)$

3) $RC = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

donde z_α satisface $P(Z < z_\alpha) = 1 - \alpha$.

Tests para la media cuando la varianza es desconocida: Supongamos ahora que la varianza es desconocida y consideremos las mismas hipótesis sobre μ:

1) H_0: $\mu = \mu_0$, H_1: $\mu < \mu_0$; (contra menor)

2) H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$; (contra mayor)

3) H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$; (bilateral)

Estadístico: $T = \sqrt{n}\frac{\bar{X} - \mu_0}{S}$

Bajo $\mu = \mu_0$ $T \sim t_{n-1}$ (t de Student con $n-1$ grados de libertad). .

Regiones de rechazo son:

1) $RC = (-\infty, -t_\alpha]$

2) $RC = [t_\alpha, \infty)$

3) $RC = (-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty)$

donde t_α satisface $P(T < z_\alpha) = 1 - \alpha$, que se encuentra en la tabla de la t de Student.

Tests para la varianza cuando la media es desconocida: Las hipótesis a testear son

1) H_0: $\sigma^2 = \sigma_0^2$, H_1: $\sigma^2 < \sigma_0^2$; (contra menor)

2) H_0: $\sigma^2 = \sigma_0^2$, H_1: $\sigma^2 > \sigma_0^2$; (contra mayor)

3) H_0: $\sigma^2 = \sigma_0^2$, H_1: $\sigma^2 \neq \sigma_0^2$; (bilateral)

Estadístico: $T = \frac{(n-1)S^2}{\sigma_0^2}$

Bajo la hipótesis H_0 ($\sigma^2 = \sigma_0^2$) el estadístico $T \sim \chi^2_{n-1}$ (Qui-cuadrado con $n-1$ grados de libertad).

Regiones de rechazo son:

1) $RC = (-\infty, -x_\alpha]$

2) $RC = [\chi^2_{1-\alpha}, \infty)$

3) $RC = (-\infty, x_{\alpha/2}] \cup [x^+_{1-\alpha/2}, \infty)$

donde x_α satisface $P(\chi^2_{n-1} < x_\alpha) = \alpha$. Esos valores se encuentran tabla de la χ^2 con $n-1$ grados de libertad.

Ejemplo Se toman 25 determinaciones de la temperatura en cierto sector de un reactor, obteniéndose

$\bar{x} = 249^oC$ y $s = 2{,}8^oC$

Interesa saber, a nivel $\alpha = 0{,}05$

a) si existe evidencia para decidir que la temperatura media en ese sector del reactor es menor que 250^oC .

b) si existe evidencia para decidir que la varianza de la temperatura en ese sector del reactor es mayor que $(2^oC)^2$.

a) Las hipótesis a testear son H_o: $\mu = 250$ (ó $\mu \geq 250$) vs H_1: $\mu < 250$.

El estadístico del test será $T = \sqrt{n}\frac{\bar{X}-\mu_0}{S}$ y la región de rechazo para ese estadístico será

$$\text{RC} = (-\infty, -t_{n-1,(0{,}05)}] = (-\infty, -1{,}71] \tag{60}$$

donde la segunda identidad se obtiene recordando que $n = 25$ y observando en la tabla que $-t_{24,(0{,}05)} = -1{,}71$.

El valor observado de T es

$$T_{\text{observado}} = 5\frac{249-250}{2{,}8} = -1,7857 \tag{61}$$

Como $-1,7857$ pertenece a la RC, se rechaza H_0. Concluimos que hay evidencia de que la temperatura media del reactor es menor que 250^oC.

b) Las hipótesis a testear son

$$H_0 : \sigma^2 = 4 \text{ (ó } \sigma^2 \leq 4$$
$$H_1 : \sigma^2 > 4$$

El estadístico del test será $T = \frac{(n-1)S^2}{\sigma_0^2}$ que tiene distribución χ^2_{n-1} y la región de rechazo quedará

$$\text{RC} = [\chi^2_{n-1,(0{,}05)}, \infty) = [36{,}42, \infty) \tag{62}$$

porque en nuestro caso, $n = 25$ y por lo tanto $\chi^2_{24,(0{,}05)} = 36{,}42$. Como el valor observado de T es

$$T_{\text{observado}} = \frac{24(2{,}8)^2}{4} = 47{,}04 \tag{63}$$

Como $47{,}04 \in$ RC, se rechaza H_0. Es decir, hay evidencia de que la varianza de la temperatura del reactor es mayor que $(2^oC)^2$.

Tests de hipótesis de nivel aproximado (o asintótico) α para la media de una distribución cualquiera: Queremos testear la media μ asumiendo la varianza σ^2 finita pero desconocida.

Usaremos el estadístico $T = \sqrt{n}\frac{\bar{X}-\mu_0}{S}$ que tiene distribución asintótica $N(0,1)$ por el TCL.

Se toma n "grande" y se trabaja como en el caso de $X \sim N(\mu, \sigma^2)$. Las regiones de rechazo son

1) $RC = (-\infty, -z_\alpha]$

2) $RC = [z_\alpha, \infty)$

3) $RC = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

donde z_α satisface $P(Z < z_\alpha) = 1 - \alpha$, $Z \sim N(0,1)$.

Test de hipótesis asintótico para p de la Bernoulli

Hay tres posibles tests para p:

1) H_0: $p = p_0$, H_1: $p < p_0$; (contra menor)

2) H_0: $p = p_0$, H_1: $p > p_0$; (contra mayor)

3) H_0: $p = p_0$, H_1: $p \neq p_0$; (bilateral)

Usamos el estadístico

$$T = \sqrt{n}\frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}},$$

Como bajo H_0, $T \sim N(0,1)$ asintoticamente (TCL), las regiones de rechazo a nivel α son, respectivamente:

1) $RC = (-\infty, -z_\alpha]$

2) $RC = [z_\alpha, \infty)$

3) $RC = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

donde z_α satisface $P(Z < z_\alpha) = 1 - \alpha$.

Ejemplo del adivino Un sujeto acierta el color de 850 de 1600 cartas puestas al dorso. Queremos decidir si creemos que es adivino.

Sea p la probabilidad que el adivino acierte. Queremos testar

$H_0 : \ p = 1/2$ (es decir, no mejora el puro azar) contra $H_1 : \ p > 1/2$ (tiene probabilidad de adivinar mayor que $1/2$).

Usando que bajo H_0 el parámetro es $p_0 = 1/2$, el estadístico observado es

$$t_{\text{obs}} = \sqrt{n}\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} = \sqrt{1600}\frac{\frac{850}{1600} - \frac{1}{2}}{\frac{1}{2}} = 2{,}5$$

que implica un P-valor $= P(T > 2{,}5) = 1 - 0{,}9938 = 0{,}0062$ (por la tabla de la normal). Es decir que podemos rechazar H_0 para cualquier $\alpha > 0{,}0062$.

Si el sujeto hubiese adivinado 825 cartas el estadístico sería

$$t_{\text{obs}} = \sqrt{1600}\frac{\frac{820}{1600} - \frac{1}{2}}{\frac{1}{2}} = 1{,}25$$

Aquí el P-valor es 0,105 que es mucho menos contundente.

Relación entre intervalos de confianza y tests bilaterales

Asumamos $X \sim N(\mu, \sigma^2)$. Sea $X_1, \ldots, X_n$ una muestra aleatoria de X.

Sabemos que el intervalo de confianza para μ de confianza $1 - \alpha$ está dado por

$$IC = \left[\bar{X} - z\frac{\sigma}{\sqrt{n}},\ \bar{X} + z\frac{\sigma}{\sqrt{n}}\right]$$

Supongamos que queremos testear las hipótesis

$$H_0 : \ \mu = \mu_0, \qquad H_1 : \ \mu \neq \mu_0$$

Si μ_0 no pertenece al intervalo de confianza, sospechamos que H_0 es falsa.

De hecho,

$$P_{\mu_0}(IC \not\ni \mu_0) = 1 - P(IC \ni \mu_0) = 1 - (1 - \alpha) = \alpha$$

O sea que rechazar H_0 si μ_0 no pertenece al intervalo de confianza $(1 - \alpha)$ nos dá un test de nivel de significancia α.

16. Tests no paramétricos

Basado en notas del Curso de Estadística del Instituto de Matemática y Estadística de la Universidad de San Pablo.

Tests de adherencia Objetivo: Testear si un modelo probabilístico es adecuado para un conjunto de datos observados.

Exemplo 1: Genética – Equilibrio de Hardy-Weinberg

Consideramos los hijos de un padre de genotipo Aa y una madre de genotipo Aa.

El **modelo teórico** dice que las probabilidades de los genotipos de los hijos son:

Tipo	AA	Aa	aa
Probab	1/4	1/2	1/4

Hay 3 categorias: AA, Aa, aa

Se estudian 100 descendientes de una pareja con esos genotipos y se observan

Genotipo	AA	Aa	aa	Total
Frecuencia observada	26	45	29	100

El objetivo es verificar si el modelo genético propuesto es adecuado para esa población.

Si el modelo es adecuado, el número esperado de descendientes para cada genotipo se calcula así:

$E_{AA} := 100\,P(AA) = 100\frac{1}{4} = 25$

$E_{Aa} := 100\,P(Aa) = 100\frac{1}{2} = 50$

$E_{aa} := 100\,P(aa) = 100\frac{1}{2} = 50$

Construimos una tabla para las frecuencias esperadas y observadas:

Genotipo	AA	Aa	aa	Total
Frecuencia observada O_i	26	45	29	100
Frecuencia esperada E_i	25	50	25	100

Podemos afirmar que los valores observados están suficientemente cerca de los esperados, de tal manera que el modelo de Hardy-Weinberg es adecuado a esta población?

Test de Adherencia – Metodología

Considere una tabla de frecuencias observadas de $k \geq 2$ categorías de resultados en n observaciones:

Categorías	1	2	...	k	Total
Frecuencia observada	O_1	O_2	...	O_k	n

donde O_i es el total de individuos observados en la categoría i, $i = 1, \ldots, k$.

Sea p_i la probabilidad teórica asociada a la categoría i.

El objetivo es testear las hipótesis

$H_0 : p_1 = p_{o1}, \ldots, p_k = p_{ok}$

H_1 : existe por lo menos una diferencia.

Aquí p_{oi} es la probabilidad asociada al modelo que estamos testeando.

Si E_i es el número esperado de individuos en la categoría i cuando H_0 es verdadera, entonces

$$E_i = np_{oi}, \quad i = 1, \ldots, k.$$

La tabla de frecuencias observadas y esperadas es

Categorías	1	2	...	k	Total
Frecuencia observada	O_1	O_2	...	O_k	n
Frecuencia esperada	E_1	E_2	...	E_k	n

Definimos el estadístico

$$\chi^2_{k-1}(\underline{O}) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

donde $\underline{O} = (O_1, \ldots, O_k)$ son funciones de la muestra aleatoria y por lo tanto variables aleatorias.

Suponiendo que H_0 es verdadera, ese estadístico tiene distribución asintótica Chi-cuadrado con $k-1$ grados de libertad. Sus probabilidades están tabuladas.

Este resultado es válido grosso modo para n grande y para valores esperados $E_i \geq 5$.

Basamos la regla de decisión en el P-valor. En ese caso,

$$P(\underline{o}) = P(\chi^2_{k-1}(\underline{O}) \geq \chi^2_{k-1}(\underline{o})),$$

Si para α fijado obtenemos $P(\underline{o}) \leq \alpha$, rechazamos H_0, si no, no rechazamos.

En el ejemplo, las hipótesis son:

H_0 : el modelo de Hardy-Weinberg es adecuado a la situación.

H_1 : el modelo no es adecuado.

Equivalentemente,

H_0: $p_0(AA) = 1/4$, $p_0(Aa) = 1/2$ e $p_0(aa) = 1/4$

H_1: por lo menos una de las tres igualdades no se verifica.

La tabla presenta los valores observados y esperados calculados antes.

Genotipo	AA	Aa	aa	Total
Frecuencia observada O_i	26	45	29	100
Frecuencia esperada E_i	25	50	25	100

Cálculo del valor de la estadística observada del test ($k = 3$):

$$\chi^2_{k-1}(\underline{o}) = 0{,}04 + 0{,}50 + 0{,}64 = 1{,}18$$

Usando la distribución de qui-cuadrado con $k - 1 = 2$ grados de libertad, el P-valor es

$$P = P(\chi^2_2 \geq 1{,}18) = 0{,}5543$$

Conclusión: Para $\alpha = 0{,}05$, como $P = 0{,}5543 > 0{,}05$, no rechazamos H_0, es decir que no hay evidencia que la población no siga el equilibrio de Hardy-Weinberg.

Tests de Independencia

Objetivo: Verificar si hay independencia entre dos variables.

Ejemplo: Queremos verificar si hay dependencia entre renta y número de hijos en las familias de una ciudad.

Son elegidas 250 familias al azar y se obtiene la tabla siguiente:

Renta \ # de hijos	0	1	2	≥ 3	Total
menos de 2000	15	27	50	43	135
2000 a 5000	25	30	12	8	75
más de 5000	8	13	9	10	40
Total	48	70	71	61	250

Los datos se refieren a dos variables aleatorias X e Y observadas en una muestra de tamaño n en forma de tabla

Hipótesis que serán testeadas

Test de independencia

H_0: X e Y son variables independientes.

H_1: X e Y no son independientes.

Cuantas observaciones debería haber en cada celda de la tabla si X e Y fueran independientes?

En ese caso las probabilidades conjuntas deberían ser iguales al producto de las probabilidades marginales:

$$p_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j)$$

y el número esperado de observaciones debería ser

$$E_{ij} = np_{ij} = np_{(i\cdot)}p_{(\cdot j)} = \frac{n_{(i\cdot)}n_{(\cdot j)}}{n}$$

bajo la hipótesis de independencia.

$n_{(i\cdot)}$:= número de observaciones de $X = i$.

$n_{(\cdot j)}$:= número de observaciones de $Y = j$.

n_{ij} := número de observaciones de $X = i$ conjunto con $Y = j$.

El estadístico propuesto bajo la suposición de independencia está dado por:

$$\chi^2_q(\underline{O}) = \sum_{i,j} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

donde $O_{ij} = n_{ij}$ representa el número total de observações en la celda (i, j).

Bajo la hipótesis de independencia $\chi^2_q(\underline{O})$ tiene distribución asintótica Chi-cuadrado de q grados de libertad.

$q := (f-1)(c-1)$, f :=número de filas; $c :=$ número de columnas.

La regla de decisión se basa en el P-valor

$$P(\underline{o}) = P(\chi^2_q(\underline{O}) \geq \chi^2_q(\underline{o}))$$

Si para α fijo obtenemos $p \geq \alpha$, rechazamos H_0, en caso contrario no podemos rechazar.

Continuación del ejemplo: renta y número de hijos. $n = 250$.

H_0 : renta y número de hijos son variables independientes.

H_1 : existe dependencia entre esas variables.

Valores esperados bajo independencia:

Renta \ # de hijos	0	1	2	≥ 3	Total
menos de 2000	25.92	37.80	38.34	32.94	135
2000 a 5000	14.40	21	21.30	18.30	75
más de 5000	7.68	11.20	11.36	9.76	40
Total	48	70	71	61	250

Donde, por ejemplo:

$$11{,}20 = \frac{70 \times 40}{250}$$

El estadístico chi-quadrado observado es

$$\chi^2_q(\underline{o}) = \ldots \text{cuentas} \cdots = 36,62$$

Determinación del número de grados de libertad:

Categorías de renta: $f = 3$

Categorías de número de hijos: $c = 4$

$$q = (f-1)(c-1) = 2 \times 3 = 6$$

El P-valor observado es $P(\underline{o}) = P(\chi^2_6 \geq 36,62) = 0,000$ (por la tabla de la χ^2_6)

Como $P = 0,000 < \alpha = 0,05$ (por ejemplo), rechazamos la independencia entre el número de hijos y la renta familiar a nivel 0,05. (Y para muchos otros valores de α menores.)

Modelos no paramétricos Basado en el Curso de modelos no paramétricos de Pedro Delicado, Universidad de Cataluña.

Modelos paramétricos versus no paramétricos X sigue un modelo paramétrico si su distribución de probabilidad F pertenece a una familia de distribuciones indexada por un parámetro θ de dimensión finita:

$$X \sim F, \quad F \in \{\mathcal{F}_\Theta = \{F_\theta,\ \theta \in \Theta \subset \mathbb{R}^k\}$$

La familia de distribuciones $\mathcal{F}_\Theta$ recibe el nombre de modelo estadístico paramétrico.

Diremos que X sigue un modelo estadístico no paramétrico si sobre su distribución F unicamente se suponen algunas condiciones de regularidad. Por ejemplo: F es una función de distribución continua.

Métodos no paramétricos Son métodos de inferencia estadística válidos cuando no se hacen hipótesis paramétricas sobre la distribución de los datos.

Test de bondad de ajuste

Sea X v.a. con función de distribución F desconocida.

Sea F_0 una función de distribución conocida. Se desea testear

$H_0 : F = F_0$

$H_1 : F \neq F_0$

También se pueden considerar las hipótesis alternativas unilaterales:

$H_1 : F(x) < F_0(x)$ para todo x

$H_1 : F(x) > F_0(x)$ para todo x

Disponemos de una muestra $\underline{X} = (X_1, \ldots, X_n)$ de X.

Vamos a estudiar el test de **Kolmogorov-Smirnov**.

Distribución empírica: Definimos $F_n = F_n(\underline{x}, x)$ por

$$F_n(\underline{x}, x) = \frac{1}{n} \sum_i \mathbf{1}\{x_i \leq x\}$$

cuenta la proporción de observaciones x_i que son menores o iguales a x. Notación: $\underline{x} = (x_1, \ldots, x_n)$.

Para $\underline{x}$ fijo, $F_n(\underline{x}, x)$ como función de x es una función de distribución: Está entre 0 y 1, es continua a la derecha, el límite a la izquierda es 0: $F_n(\underline{x}, x)$ es cero para $x < \min x_i$ y el límite a la derecha es 1: $F_n(\underline{x}, x)$ es cero para $x > \max x_i$ y es no claramente no decreciente.

Como $F_n(\underline{x}, \cdot)$ depende de $\underline{x}$, $F_n(\underline{X}, \cdot)$ es una función del vector aleatorio $\underline{X}$ y por lo tanto es una función de distribución aleatoria.

Para cada x fijo cada término $\mathbf{1}\{X_i \leq x\}$ es una variable aleatoria de Bernoulli con probabilidad de éxito $p = P(\mathbf{1}\{X_i \leq x\} = 1) = P(X_i \leq x) = F(x)$

Escribimos $F_n(x)$ en lugar de $F_n(\underline{X}, x)$.

$F_n(x)$ es una variable aleatoria y $nF_n(x)$ tiene distribución binomial con parámetros n y $p = F(x)$.

Propiedades

1) $EF_n(x) = F(x)$ para cada $x \in \mathbb{R}$.

2) Por la ley de grandes números $\lim_{n\to\infty} F_n(x) = F(x)$ en probabilidad, para cada $x \in \mathbb{R}$.

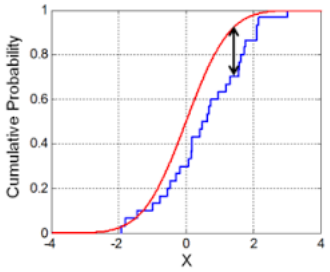
3) Por el Teorema Central del Límite,

$$\lim_{n\to\infty} \sqrt{n} \frac{F_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} = Z \qquad \text{en distribución}$$

donde $Z \sim N(0, 1)$.

Definición

$$D_n^+ := \sup_{x\in\mathbb{R}}(F_n(x) - F(x)), \quad D_n^- := \sup_{x\in\mathbb{R}}(F(x) - F_n(x))$$



$$D_n := \text{máx}\{D_n^+, D_n^-\} = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

4) Teorema de Glivenko Cantelli.

$$\lim_{n\to\infty} D_n = 0$$

Esto no lo probaremos.

5) Se pueden demostrar las siguientes convergencias en distribución. Para $z > 0$,

$$\lim_{n\to\infty} P(\sqrt{n}D_n^{\pm} > z) = e^{-2z^2}$$

$$\lim_{n\to\infty} P(\sqrt{n}D_n > z) = 2\sum_{i=0}^{\infty}(-1)^{i-1}e^{-2i^2z^2}$$

6) Para n "grande"

$$4n(D_n^+)^2 \sim \chi_2^2$$

Es decir que el supremo de la diferencia converge a una distribución chi-cuadrado de 2 grados de libertad.

Vamos a establecer la región crítica y el P-valor para los tres tests de bondad de ajuste

H_0	H_1	RC (α)	P-valor
$F = F_0$	$F \neq F_0$	$D_n(\underline{x}) \geq d_{n,\alpha}$	$P(D_n \geq D_n(\underline{x}))$
$F = F_0$	$F > F_0$	$D_n^+(\underline{x}) \geq d_{n,\alpha}^+$	$P(D_n^+ \geq D_n^+(\underline{x}))$
$F = F_0$	$F < F_0$	$D_n^-(\underline{x}) \geq d_{n,\alpha}^-$	$P(D_n^- \geq D_n^-(\underline{x}))$

donde $D_n(\underline{x})$ son los valores observados, $d_{n,\alpha}$ está definido por $P(D_n > d_{n,\alpha}) = \alpha$, etc.

Ejemplo Queremos saber si los valores $\{1; 7; 2; 5; 5{,}3\}$ vienen de una distribución mayor que la uniforme en $[0, 10]$.

$H_0 : F(x) = F_0(x) = \frac{x}{10}$ en $[0, 10]$, etc.

$H_1 : F(x) > F_0(x)$.

Ordenamos los datos: 1;2;5;5.3;7

Calculamos la distribución empírica:

F_n	$F_n - F$	intervalo
0	0	$x < 0$
0	$-\frac{x}{10}$	$0 \leq x < 1$
$\frac{1}{5}$	$\frac{1}{5} - \frac{x}{10}$	$1 \leq x < 2$
$\frac{2}{5}$	$\frac{2}{5} - \frac{x}{10}$	$2 \leq x < 5$
$\frac{3}{5}$	$\frac{3}{5} - \frac{x}{10}$	$5 \leq x < 5{,}3$
$\frac{4}{5}$	$\frac{4}{5} - \frac{x}{10}$	$5{,}3 \leq x < 7$
1	$1 - \frac{x}{10}$	$7 \leq x < 10$
1	0	$10 \leq x$

de donde $d_n^+(\underline{x}) = \sup_x F_n(x) - F(x) = \frac{3}{10}$.

$4n(d_n^+(\underline{x}))^2 = 4 \times 5 \times \frac{9}{100} = 1{,}8$

P-valor $= P(\chi_2^2 > 1{,}8) = 0{,}4$. No se puede rechazar H_0.

Kolmogorov Smirnov para dos muestras Queremos testear si dos muestras de tamaños n y m $X_1, \ldots, X_n$ de X y $Y_1, \ldots, Y_m$ de Y vienen de la misma distribución.

$H_0 : F_X = F_Y$

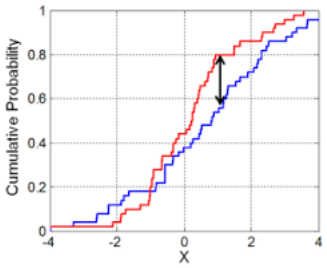
$H_1 : F_X(x) \neq F_Y(x)$ para todo x.

Usamos las distribuciones empíricas

$$F_n(\underline{x}, x) = \frac{1}{n} \sum_i \mathbf{1}\{x_i \leq x\}$$
$$F_m(\underline{y}, y) = \frac{1}{m} \sum_i \mathbf{1}\{y_i \leq y\}$$

y definimos el estadístico observado

$$D_{n,m}(\underline{x}, \underline{y}) := \sup_x |F_n(\underline{x}, x) - F_m(\underline{y}, x)| \tag{64}$$



La hipótesis nula se rechaza si

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}} \tag{65}$$

donde $c(\alpha)$ fué calculado para algunos valores de α:

α	0,10	0,05	0,025	0,01	0,005	0,001
$c(\alpha)$	1,22	1,36	1,48	1,63	1,73	1,95

Al rechazar la hipótesis este test dice que las dos muestras no vienen de la misma distribución. En caso de no rechazarle, dice que no se puede rechazar que vengan de la misma distribución pero no especifica de cual distribución vendrían.

Las tablas de los valores críticos están publicadas.

Test de rango para dos muestras

Queremos testear si dos muestras del mismo tamaño $X_1, \ldots, X_n$ de X y $Y_1, \ldots, Y_n$ de Y vienen de la misma distribución.

$H_0 : F_X = F_Y$

$H_1 : F_X(x) > F_Y(x)$ para todo x.

Supongamos F_X continua. Todas las observaciones son distintas.

Para construir el estadístico, primero ordenamos las muestras. Definiendo

$A = \{X_1, \ldots, X_n, Y_1, \ldots, Y_n\}$

$$T_k = \text{mín}(A \setminus \{T_1, \ldots, T_{k-1}\}), \quad k = 1, \ldots, 2n.$$

Y construimos la trayectoria de un paseo aleatorio: $S_0 = 0$

$$S_k = S_{k-1} + \mathbf{1}\{T_k \in \underline{X}\} - \mathbf{1}\{T_k \in \underline{Y}\}$$

Vamos recorriendo las observaciones ordenadas y subiendo 1 cuando la observación viene de la muestra X y bajando 1 cuando viene de la muestra Y.

Como el tamaño de las muestras es el mismo, el paseo aleatorio termina en 0 en el instante $2n$.

Bajo la hipótesis H_0 todas las combinaciones de subidas y bajadas tienen la misma probabilidad $1/2^n$ y el máximo

$$M_{2n} := \text{máx}\{S_k,\, k = 0, \ldots, 2n\}$$

del paseo aleatorio S_n satisface el siguiente límite asintótico:

$$\lim_{n\to\infty} P\Big(\frac{M_{2n}}{\sqrt{2n}} \geq b \;\Big|\; S_{2n} = 0\Big) = e^{-2b^2}$$

Por otra parte, asintoticamente,

$$\frac{M_{2n}^2}{2n} \sim \chi_2^2$$

Con esto en manos podemos construir nuestro test.

17. Correlación y Regresión lineal

Basado en MIT OpenCourseWare http://ocw.mit.edu 15.075J / ESD.07J Statistical Thinking and Data Analysis Fall 2011

Objetivo: expresar la intensidad de la relación entre dos variables.

Usaremos:

Análisis de correlación, que consiste en medir la intensidad y dirección de la relación lineal entre dos variables.

Regresión. Predecir o estimar los resultados de una variable basados en otra variable o variables.

Correlación. Graficar las observaciónes de las dos variables en el plano.

Inspeccionar visualmente el gráfico. Hay una relación lineal? Hay outliers?

Coeficiente de correlación Mide la intensidad y dirección entre dos variables X e Y.

Coeficiente de correlación poblacional:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{VX\,VY}} \tag{66}$$

toma valores en $[0, 1]$.

$\rho = -1$ correlación lineal negativa perfecta.

$\rho = 1$ correlación lineal positiva perfecta.

$\rho = 0$: no hay correlación lineal.

Coeficiente de correlación muestral:

$$r = \frac{\text{Cov Muestral}}{\sqrt{s_X^2 s_Y^2}} = \frac{\frac{\sum_i (x_i - \bar{x})(x_i - \bar{x})}{n-1}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}} \tag{67}$$

Conviene standarizar las variables para calcular r:

$$\frac{x_i - \bar{x}}{\sqrt{s_X^2}}, \frac{y_i - \bar{y}}{\sqrt{s_Y^2}} \tag{68}$$

Si se observa una elipse, la correlación es la inclinación del eje mayor.

Otros nombres que se le dan a r son: Coeficiente de correlación de Pearson, momento producto de correlación

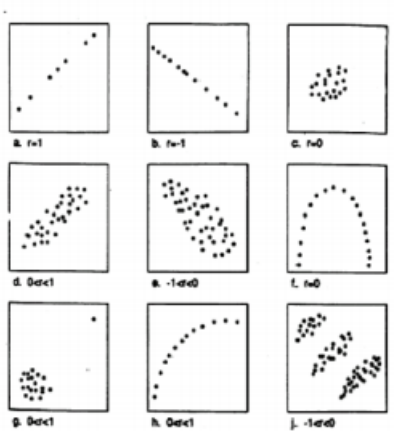
r mide la *correlación lineal* entre las variables.

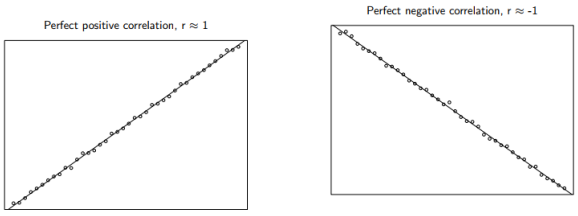
El valor de r es independiente de las unidades en que están expresadas las variables.

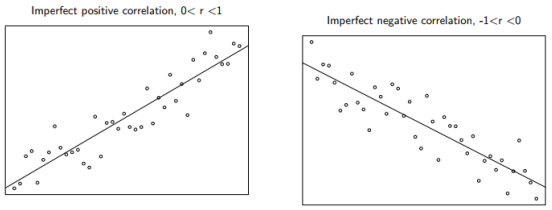
r es sensible a outliers.

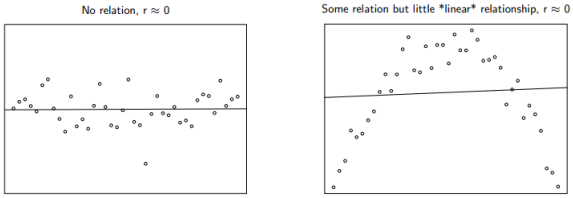
r^2 nos dice la proporción de la variación en Y que puede ser explicada por su relación lineal con X

Ejemplos:





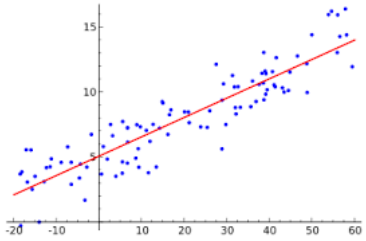




Regresión lineal simple

La regresión permite estimar la relación de una variable respuesta Y a una variable X llamada predictora. También se puede trabajar con X e Y vectores, pero aquí nos concentraremos en el caso unidimensional.

Sean $x_1, \ldots, x_n$ valores de la variable X elegidos por el investigador.



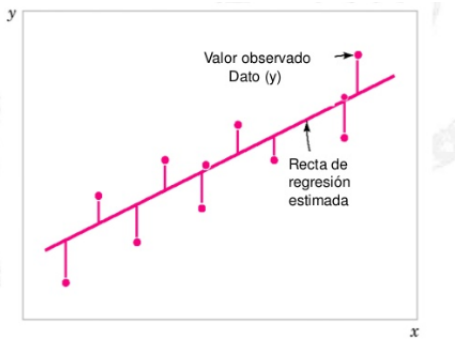
Los valores $y_1, \ldots, y_n$ son una muestra de variables $Y_1, \ldots, Y_n$, tales que la distribución de Y_i depende del valor no aleatorio x_i de la siguiente manera

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{69}$$

donde $\varepsilon_i \sim N(0, \sigma^2)$. Así, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ No hace falta suponer que los ε_i son normales para hallar la recta de mínimos cuadrados. Eso se necesita para hacer inferencia.

Para encontrar la recta estimada, buscamos los valores de β_0 y β_1 que minimizan la suma de los cuadrados:

$$Q := \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{70}$$



Para encontrarlos, derivamos, igualamos a 0 y llamamos $\hat{\beta}_0$ y $\hat{\beta}_1$ la solución de

$$0 = \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) \tag{71}$$

$$0 = \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - (\beta_0 + \beta_1 x_i)) \tag{72}$$

Reescribiendo (71),

$$\begin{aligned}
&\sum_i y_i - \sum_i \hat{\beta}_0 - \sum_i \hat{\beta}_1 x_i = 0 \\
&\sum_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i = 0 \\
&\frac{1}{n} \sum_i y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_i x_i = 0 \\
&\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \\
&\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned} \tag{73}$$

Resolviendo la ecuación (72):

$$\begin{aligned}
&\sum_i x_i y_i - \sum_i x_i \hat{\beta}_0 - \sum_i \hat{\beta}_1 x_i^2 = 0 \\
&\sum_i x_i y_i - \hat{\beta}_0 \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 = 0 \\
&\sum_i x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 = 0 \\
&\sum_i x_i y_i - \bar{y} \sum_i x_i - \hat{\beta}_1 \frac{1}{n} \Big(\sum_i x_i \Big)^2 - \hat{\beta}_1 \sum_i x_i^2 = 0
\end{aligned}$$

De donde

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i}{\sum_i x_i^2 - \frac{1}{n} \Big(\sum_i x_i \Big)^2} \tag{74}$$

Defina

$$\begin{aligned}
\tilde{s}_{xy} &:= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \frac{1}{n} \sum_i x_i y_i \\
\tilde{s}_{xx} &:= \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - \frac{1}{n} \sum_i x_i^2
\end{aligned}$$

Así:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\tilde{s}_{xy}}{\tilde{s}_{xx}} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned} \tag{75}$$

Procedimiento para encontrar la recta de mínimos cuadrados Dados $x_1, \ldots, x_n$ y $y_1, \ldots, y_n$ calcule $\bar{x}$, $\bar{y}$, $\tilde{s}_{x,y}$, $\tilde{s}_{x,x}$ y calcule

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\tilde{s}_{xy}}{\tilde{s}_{xx}} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned} \tag{76}$$

y la respuesta es

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0. \tag{77}$$

es la recta estimada que está mas cerca de los puntos en el sentido de mínimos cuadrados. Graficar.

Bondad de ajuste Defina

$$\begin{aligned}
\text{SSE} &= \text{suma de los errores cuadráticos en relación a la recta estimada} \\
&= \sum_i (y_i - \hat{y}_i)^2
\end{aligned}$$

La suma total de los cuadrados mide la variación cuadrática de y_i alrededor de su media $\bar{y}$

$$\begin{aligned}
\text{SST} &= \text{suma total de los errores cuadráticos en relación a } \bar{y} \\
&= \sum_i (y_i - \bar{y})^2 =: \tilde{s}_{yy}
\end{aligned}$$

De donde, se puede demostrar que

$$\begin{aligned}
\text{SST} &= \sum_i (y_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 = \text{SSE} + \text{SSR}
\end{aligned} \tag{78}$$

donde SSR es la suma de los cuadrados de las regresiones. Es la variación cuadrática del modelo en relación a su media muestral.

Considere

$$r^2 := \frac{\text{SSR}}{\text{SST}} = \frac{\text{variación del modelo}}{\text{variación total}} =: \text{coeficiente de determinación} \tag{79}$$

Resulta que r^2 es el cuadrado del coeficiente de correlación muestral

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} \tag{80}$$

Veamos eso:

$$\begin{aligned}\text{SSR} &= \sum_i (\hat{y}_i - \bar{y})^2 \\ &= \sum_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})\right)^2 \\ &= \hat{\beta}_1^2 \sum_i \left(x_i - \bar{x}\right)^2 = \hat{\beta}_1^2 \tilde{s}_{xx}\end{aligned}$$

Ahora veamos r^2:

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1^2 \tilde{s}_{xx}}{\tilde{s}_{yy}} = \frac{s_{xy}^2 \tilde{s}_{xx}}{\tilde{s}_{xx}^2 \tilde{s}_{yy}} = \frac{s_{xy}^2}{\tilde{s}_{xx}\tilde{s}_{yy}} \tag{81}$$

o sea que r^2 es el cuadrado del coeficiente de correlación muestral. Indica cuánto de la variación total es explicado por la regresión lineal.

La varianza σ^2 mide la dispersión de los y_i alrededor de las medias $\mu_i = \beta_0 + \beta_1 x_i$. Un estimador insesgado de σ^2 es

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2} \tag{82}$$

Alternativamente podemos definir

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

si la regresion funciona perfectamente, entonces $R^2 = 1$, y si no funciona nada entonces $R^2 = 0$ (en este caso la mejor prediccion es la media para todos los $\hat{y}_i$).

Y despues se demuestra que $R^2 = r^2$ y que entonces, da un estimador de la correlacion ρ con la que empezamos.

Inferencia. Queremos estimar β_0, β_1. Recordemos que

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{83}$$

donde los errores son Normal$(0, \sigma^2)$. Se puede demostrar que $\hat{\beta}_0$ y $\hat{\beta}_1$ son normales con

$$\begin{aligned} E\hat{\beta}_0 &= \beta_0, \quad V\hat{\beta}_0 = \sigma^2 \frac{\sum_i x_i^2}{n\tilde{s}_{xx}} \\ E\hat{\beta}_1 &= \beta_1, \quad V\hat{\beta}_1 = \frac{\sigma^2}{\tilde{s}_{xx}} \end{aligned}$$

Si definimos

$$S^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2} \tag{84}$$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ es una función de $\underline{Y}$ a traves del procedimiento de mínimos cuadrados. Se puede demostrar que

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}. \tag{85}$$

Podemos construir intervalos de confianza y hacer tests sobre β_0, β_1 usando $\hat{\beta}_0, \hat{\beta}_1$ como estimadores de sus esperanzas β_0, β_1. Tenemos que

$$S_{\hat{\beta}_0} = S\sqrt{\frac{\sum_i x_i^2}{n\tilde{s}_{xx}}}, \quad S_{\hat{\beta}_1} = \frac{S}{\sqrt{\tilde{s}_{xx}}} \tag{86}$$

son estimadores de $\sqrt{V\hat{\beta}_0}$ y $\sqrt{V\hat{\beta}_1}$, respectivamente.

Además se puede ver (cociente de normal y chi cuadrado independientes) que

$$\frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \sim t_{n-2}, \qquad \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t_{n-2}. \tag{87}$$

Así, los intervalos de confianza $(1-\alpha)$ están dados por

$$\begin{aligned} &\hat{\beta}_0 \pm t_{n-2,\alpha/2} S_{\hat{\beta}_0}, \\ &\hat{\beta}_1 \pm t_{n-2,\alpha/2} S_{\hat{\beta}_1} \end{aligned} \tag{88}$$

Test de hipótesis para β_1:

$$\begin{aligned} H_0 &: \beta = \beta_1^0 \\ H_0 &: \beta \neq \beta_1^0 \end{aligned}$$

y se rechaza H_0 a nivel α si

$$\frac{|\hat{\beta}_{1,\text{observ}} - \beta_1^0|}{S_{\hat{\beta}_1}} > t_{n-2,\alpha/2} \tag{89}$$

Si se testea $\beta_1 = 0$, estamos testeando si hay una relación linear entre x e y. Si se rechaza $\beta_1 = 0$, concluimos que y depende de x.

Cuando $\beta_1 = 0$, $t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$

18. Aplicaciones

18.1. El problema "Coupon collector" o álbum de figuritas

El coleccionador de figuritas quiere completar su album con n figuritas distintas. Para simplificar suponemos que las figuritas aparecen de a una, elegida uniformemente en $\{1, \ldots, n\}$ y que son independientes. Es decir sorteamos sucesivamente, **con reposición**, n objetos (o figuritas) distintos. Este modelo se usa en muchos contextos: por ejemplo en un sistema peer-to peer, interesa saber cuántas conexiones hacen falta para tener todos los pedazos de un archivo.

Sean $U_1, U_2, \ldots$ variables aleatorias independientes con distribución uniforme en $\{1, \ldots, n\}$. U_i representa la i-ésima figurita que exploramos.

Sea S_i el primer instante en que nuestra colección tiene i figuritas distintas. Esa variable está definida inductivamente por $S_1 = 1$ y

$$S_i := \text{mín}\{k > S_{i-1} : U_k \neq U_i, \text{ para todo } i < k\},$$

$T = S_n$ es el tiempo para completar la colección y $T_i = S_i - S_{i-1}$ el número de figuritas exploradas entre el instante en que incorporamos la $(i-1)$-ésima y el instante en que incorporamos la i-ésima nueva figurita a la colección.

Cálculo de esperanza y varianza de T La probabilidad de tener un nuevo objeto cuando hay $i-1$

objetos en la colección es $p_i = \frac{n-(i-1)}{n}$. Entonces cada t_i es geometrica con promedio $\frac{1}{p_i}$. Entonces:

$$\begin{aligned} E(T) &= E(T_1) + E(T_2) + \cdots + E(T_n) = \frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_n} \\ &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \\ &= n\left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}\right) = nH_n, \end{aligned} \tag{90}$$

Con $H_n = \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}\right)$. Aproximando por integrales, se puede mostrar que

$$H_n = \log n + \gamma + \frac{1}{2n} + O(1/n),$$

con $\gamma \approx 0{,}5772156649$ es la constante de Euler–Mascheroni. Ver, por ejemplo eventuallyalmosteverywhere.wordpress.com/2 coupon-collector-problem/

Usando la desigualdad de Markov obtenemos $\mathcal{P}(T \geq cnH_n) \leq \frac{1}{c}$.

Ahora usando la independencia de los sorteos, la varianza es:

$$\begin{aligned} \mathrm{Var}(T) &= \mathrm{Var}(T_1) + \mathrm{Var}(T_2) + \cdots + \mathrm{Var}(T_n) \\ &= \frac{1-p_1}{p_1^2} + \frac{1-p_2}{p_2^2} + \cdots + \frac{1-p_n}{p_n^2} \\ &\leq \left(\frac{n^2}{n^2} + \frac{n^2}{(n-1)^2} + \cdots + \frac{n^2}{1^2}\right) \\ &= n^2\left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2}\right) \\ &\leq \frac{\pi^2}{6}n^2. \end{aligned} \tag{91}$$

En particular para $x > 0$, usando Chevichev,

$$\mathcal{P}(T \geq nH_n + xn) \leq \mathrm{Var}(T)\frac{1}{n^2x^2} \leq \frac{\pi^2}{6x^2}.$$

18.2. Ejemplo de Machine Learning: Restricted Boltzman machines

Basado en "An Introduction to Restricted Boltzmann Machines" por Asja Fischer and Christian Igel .

Es un ejemplo tipico de aprendizaje que puede ser eficiente en problemas de clasificación y recomendación. Este método fue utilizado como herramienta para recomendar peliculas en la competición de Netflix, pero se puede usar en muchos otras aplicaciones. En ese ejemplo, a cada usuario le corresponde un conjunto de variables personales como la edad, sexo, país, etc y sus ratings de algunas peliculas. Esta data es conocida y se denomina "visible". Con estas variables conocidas y todo lo que se conoce de los otros usuarios se quiere adivinar los ratings que este usuario le daría a cada una de las peliculas en la base. Este tipo de recomendación se llama "collaborative filtering" porque se basa en el comportamiento (personal data y ratings) de otros usuarios para adivinar los rating de uno particular. Otro metodo llamado "content-based" usa la cercanía de contenidos y no de usuarios.

La idea de RBM es de definir variables "ocultas" (hidden) que son explicativas. Estas variables no tienen interpretación a priori pero pueden a posteriori explicar un comportamiento. Por ejemplo, los adolescentes masculinos que viven en CABA y dan alto rating a Star Wars van a tener ratings muy parecidas para otras peliculas. El algoritmo puede llegar a definir una variable de esta índole, y de alguna manera permite agrupar los usuarios en grupos homogeneos a traves de estas variables ocultas. El unico input del algoritmo es el numero d de variables ocultas.

Vamos a asumir que las variables son todas binarias. Si no lo fueran siempre se pueden hacer transformaciones usando más variables para que lo sean, pero puede ser costoso en términos de memoria.

Cada usuario tiene un vector aleatorio (V, H) con $V \in \{0,1\}^m$ y $H \in \{0,1\}^d$, donde $V = (V_1, \ldots, V_m)$ son las variables visibles y $H = (H_1, \ldots, H_d)$ las variables escondidas (hidden). Este vector tiene una

distribución de tipo Gibbs, dada por

$$\mathcal{P}((V,H)=(v,h))=\frac{1}{Z}\exp(-\mathcal{E}(v,h)),$$

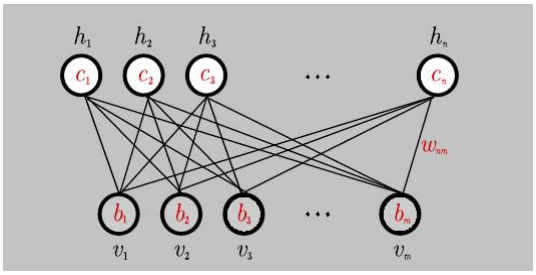
Donde Z es la constante de normalización

$$Z=\sum_{v,h}\exp(-\mathcal{E}(v,h)) \tag{92}$$

y $\mathcal{E}(v,h)$ es una medida de interacción entre las variables dada por

$$\mathcal{E}(v,h)=-\sum_{i\leq m}\sum_{j\leq d}w_{ij}v_ih_j-\sum_{i\leq m}b_iv_i-\sum_{j\leq d}c_jh_j$$

Los pesos b_i y c_i son los parámetros de sesgo y los w_{ij} son los parámetros de interacción entre las variables visibles y las ocultas. Si w_{ij} son grandes, $\mathcal{E}(v,h)$ es muy negativa, y la configuración correspondiente tiene probabilidad alta.



El algoritmo tiene dos fases: una fase de aprendizaje que usa los datos para estimar los parámetros y otra que usa lo aprendido (la estimación de los parámetros) para predecir, clasificar y recomendar.

Se supone que tenemos muchos usuarios con las variables V conocidas (o en la practica una parte significativa de V) y que vamos a aprender los pesos con esta data. Luego, para cualquier usario nuevo, podemos predecir algunos valores $V_i, i\in A$ sabiendo algunas $V_j, j\in B$. Por ejemplo, sabiendo que un usuario es adolescente, vive en Mendoza y puso un rating de 5 a House of Cards se puede inferir el rating de Game of Thrones usando los pesos w, b, c. Damos mas detalles a continuación.

Fase de aprendizaje

En la fase de aprendizaje estimaremos los parametros $\theta=(w,b,c)$ usando máxima (log)-verosimilitud. Como no se puede maximizar de manera exacta, se hace numericamente. Hay muchos algoritmos para hacer eso y hacerlo bien es el cuello de botella del método. (Stochastic gradient descent, Gibbs sampling, ...)

Fase de predicción

Una vez que estimamos los pesos, se pueden inferir los valores de rating u otros datos desconocidos de la siguiente manera. Uno primero calcula los valores de la variable H dado los conocidos de V. Se suponen que los valores desconocidos valen a priori 0.

Se puede probar (usando probabilidades condicionales y la ley de (V,H)) que:

$$\mathcal{P}(H_i=1|V=v)=\sigma\Big(\sum_j w_{ij}v_j+c_i\Big),$$

donde

$$\sigma(x)=\frac{1}{1+e^{-x}}$$

es la función sigmoid. Así, se puede decidir poner 1 en la variable H_i si su probabilidad es suficientement alta.

Una vez calculado los H, uno puede volver atrás y proponer nuevos valores de V (para los valores desconocidos), de la misma manera:

$$\mathcal{P}(V_j=1|H=h)=\sigma\Big(\sum_i w_{ij}h_i+b_j\Big),$$

se puede proponer un valor para cualquier V_i desconocido, basado en la estimación de esta probabilidad.

Expertos independientes Dada la falta de dependencia entre las variables ocultas, se puede demostrar que la distribución marginal de V tiene forma producto:

$$\mathcal{P}(V=v) = \prod_{i\leq m} e^{b_j v_j} \prod_{i\leq d} \Big(1 + e^{\sum_{j\leq m} w_{ij} v_j}\Big).$$

Entonces las variables ocultas actuan como d expertos independientes, y sus recomendaciones son combinadas de manera multiplicativa.

Maxima verosimilitud y aprendizaje

$$\begin{aligned}\log p(\theta|v) &= -\log(Z) + \log\Big(\sum_h \exp(-\mathcal{E}(v,h))\Big),\\ &= -\log\Big(\sum_{v,h} \exp(-\mathcal{E}(v,h))\Big) + \log\Big(\sum_h \exp(-\mathcal{E}(v,h))\Big).\end{aligned}$$

Eso permite escribir

$$\begin{aligned}&\frac{\partial}{\partial\theta}\log p(\theta|v)\\ &= \frac{\sum_{v,h}\exp(-\mathcal{E}(v,h))\frac{\partial}{\partial\theta}\mathcal{E}(v,h)}{\sum_{v,h}\exp(-\mathcal{E}(v,h))} - \frac{\sum_h \exp(-\mathcal{E}(v,h))\frac{\partial}{\partial\theta}\mathcal{E}(v,h)}{\sum_h \exp(-\mathcal{E}(v,h))}.\end{aligned}$$

Ahora usando que

$$p(v|h) = \frac{p(v,h)}{p(h)} = \frac{\exp(-\mathcal{E}(v,h))}{\sum_h \exp(-\mathcal{E}(v,h))},$$

obtenemos:

$$\frac{\partial}{\partial\theta}\log p(\theta|v) = \sum_{v,h} p(v,h)\frac{\partial}{\partial\theta}\mathcal{E}(v,h) - \sum_h p(h|v)\frac{\partial}{\partial\theta}\mathcal{E}(v,h).$$

Usando la definición de E,

$$\sum_h p(h|v)\frac{\partial}{\partial w_{ij}}\mathcal{E}(v,h) = \sum_h p(h|v)h_i v_j.$$

y la forma producto de la distribución:

$$\begin{aligned}\sum_h p(h|v)h_i v_j &= \sum_{h_i,h_k,k\neq i} h_i v_j p(v|h_i)\prod_{k\neq i} p(v|h_k),\\ &= \sum_{h_i} h_i v_j p(v|h_i)\sum_{h_k}\prod_{k\neq i} p(v|h_k),\\ &= P(H_i=1|v_j)\end{aligned}$$

Entonces:

$$\begin{aligned}\frac{\partial}{\partial w_{ij}}\log p(\theta|v) &= \sum_{v,h} p(h|v)p(v)\frac{\partial}{\partial w_{ij}}\mathcal{E}(v,h) - \sum_h p(h|v)\frac{\partial}{\partial w_{ij}}\mathcal{E}(v,h),\\ &= -\sum_v p(v)\mathcal{P}(H_i=1|v_j)v_j + \mathcal{P}(H_i=1|v_j)v_j.\end{aligned}$$

Para datos $(v^1, v^2, \ldots, v^n)$, y denotando $q(v)$ la distribución empirica de v, llegamos a

$$\begin{aligned}\frac{1}{n}\sum_{l\leq n}\frac{\partial}{\partial w_{ij}}\log p(\theta|v) &= \frac{1}{n}\sum_l \mathbb{E}_{p(h|v^l)}h_i v_j^l - \frac{1}{n}\sum_l \mathbb{E}_{p(v^l,h)}h_i v_j^l,\\ &= \mathbb{E}_{p(h|v)q(v)}h_i v_j - \mathbb{E}_{p(v,h)}h_i v_j,\\ &=< h_i, v_j >_{data} - < h_i, v_j >_{modelo}.\end{aligned}$$

Entonces el gradiente de la log-verosimulitud tiene una interpretación simple. Es la diferencia entre la probabilidad empirica de tener $H_i = 1$ cuando $v_i = 1$ con los pesos actuales y la data v y la probabilidad de tener $H_i = 1$ para el modelo, i.e. usando las probabilidades actuales de los v.

Es importante notar que el segundo termino es dificil calcular en general, porque involucra una suma sobre todos los v. Para aproximarlo, se usa la ley de grandes numeros para procesos de Markov (y una esperanza que haya convergencia rapida) o se usan aproximaciones aún mas brutales.