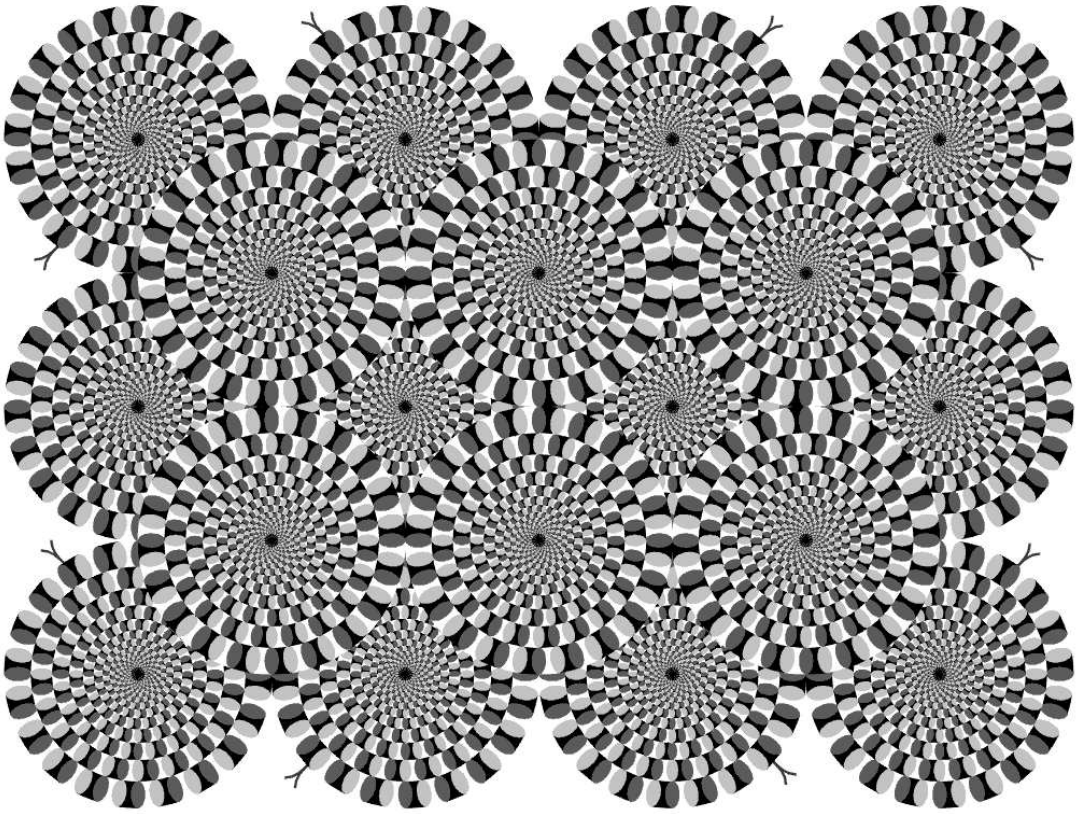
Probabilidad y Estadística (Borradores, Curso 23)
Condicionales

Sebastian Grynberg

18 de abril de 2011



Serpientes de Akiyoshi Kitaoka.

Si no se espera,
no se encontrará lo inesperado,
pues el sendero que a ello conduce
es inaccesible

(Heráclito.)

Índice

1. Condicionales

1.1. Caso discreto

Sean X e Y dos variables aleatorias discretas definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Fijemos un valor $x \in \mathbb{R}$ tal que $p_X(x) > 0$. Usando la noción de probabilidad condicional podemos definir la *función de probabilidad condicional de Y dado que $X = x$*, mediante

$$p_{Y|X}(y|x) := \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)}. \tag{1}$$

Función de distribución condicional de Y dado que $X = x$. La *función de distribución condicional de Y dado que $X = x$* se define por

$$F_{Y|X}(y\,|x) \;\; := \;\; \mathbb{P}(Y \leq y|X = x) = \sum_{t \leq y} \mathbb{P}(Y = t|X = x) = \sum_{t \leq y} p_{Y|X}(t|x). \tag{2}$$

Esperanza condicional de Y dado que $X = x$. La *esperanza condicional de Y dado que $X = x$* se define por

$$\mathbb{E}[Y|X = x] := \sum_y y\, p_{Y|X}(y|x). \tag{3}$$

Nota Bene 1. Si se quieren calcular las funciones de probabilidad de las variables $Y|X = x$, $x \in Sop(p_X)$, la fórmula (1) dice que basta dividir cada fila de la representación matricial de la función de probabilidad conjunta de X e Y, $p_{X,Y}(x, y)$ por el correspondiente valor de su margen derecho, $p_X(x)$. En la fila x de la matriz resultante se encuentra la función de probabilidad condicional de Y dado que $X = x$, $p_{Y|X}(y|x)$.

Nota Bene 2. La función $F_{Y|X}(\cdot\,|x) : \mathbb{R} \to \mathbb{R}$ definida en (2) es una *función de distribución genuina*: es no decreciente, continua a derecha, tiende a 0 cuando $y \to -\infty$ y tiende a 1 cuando $y \to \infty$. Por lo tanto, podemos interpretarla como la función de distribución de una nueva variable aleatoria, $Y|X = x$, cuya ley de distribución coincide con la de Y cuando se sabe que ocurrió el evento $\{X = x\}$. Motivo por el cual la llamaremos Y *condicional a que* $X = x$

Nota Bene 3. Todas las nociones asociadas a las distribuciones condicionales se definen de la misma manera que en el caso de una única variable aleatoria discreta, salvo que ahora todas las probabilidades se determinan condicionales al evento $X = x$. Las definiciones tienen sentido siempre y cuando $x \in Sop(p_X)$.

Ejemplo 1.1. En una urna hay 3 bolas rojas, 2 amarillas y 1 verde. Se extraen dos. Sean X e Y la cantidad de bolas rojas y amarillas extraídas, respectivamente. La representación matricial de la función de probabilidad conjunta $p_{X,Y}(x, y)$ y de sus marginales $p_X(x)$, $p_Y(y)$ es la siguiente

$X \setminus Y$	0	1	2	p_X
0	0	2/15	1/15	3/15
1	3/15	6/15	0	9/15
2	3/15	0	0	3/15
p_Y	6/15	8/15	1/15	

Cuadro 1: Distribución conjunta de X e Y y sus respectivas marginales.

Dividiendo cada fila de la matriz $p_{X,Y}(x, y)$ por el correspondiente valor de su margen derecho se obtiene el siguiente Cuadro que contiene toda la información sobre las funciones de probabilidad de las condicionales $Y|X = x$. Por ejemplo, la función de probabilidad

$X \setminus Y$	0	1	2
0	0	2/3	1/3
1	1/3	2/3	0
2	1	0	0

Cuadro 2: Distribuciones de las variables condicionales Y dado que $X = x$. Interpretación intuitiva de los resultados: *a medida que X aumenta el grado de indeterminación de Y disminuye.*

condicional de Y dado que $X = 0$, es la función de y definida en su primera fila:

$$p_{Y|X}(0|0) = 0; \qquad p_{Y|X}(1|0) = 2/3; \qquad p_{Y|X}(2|0) = 1/3.$$

Notar que la función de probabilidad condicional obtenida es diferente de la correspondiente a la marginal de Y, $p_Y(y)$. Del Cuadro 2 y la definición (3) se deduce que

$$\mathbb{E}[Y|X = x] = \frac{4}{3}\mathbf{1}\{x = 0\} + \frac{2}{3}\mathbf{1}\{x = 1\} \tag{4}$$

□

Nota Bene. Observar que en general la función de probabilidad condicional $p_{Y|X}(y|x)$ es diferente de la función de probabilidad $p_Y(y)$. Esto indica que se pueden hacer inferencias sobre los valores posibles de Y a partir de los valores observados de X y viceversa; las dos variables son (estocásticamente) *dependientes*. Más adelante veremos algunas maneras de hacer este tipo de inferencias.

1.2. Mezclas

Definición 1.2 (Mezcla)**.** Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Sea $M : \Omega \to \mathbb{R}$ una variable aleatoria discreta tal que $M(\Omega) = \mathcal{M}$ y $p_M(m) = \mathbb{P}(M = m) > 0$ para todo $m \in \mathcal{M}$. Sea $(X_m : m \in \mathcal{M})$ una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ e independiente de M. En tal caso, la variable aleatoria $X := X_M$ está bien definida y se llama la *mezcla* de las variables X_m obtenida mediante la variable *mezcladora* M.

Nota Bene. La distribución de probabilidades de M indica la proporción en que deben mezclarse las variables X_m: para cada $m \in \mathcal{M}$, la probabilidad $p_M(m)$ representa la proporción con que la variable X_m participa de la mezcla X_M. □

Cálculo de la función de distribución. La función de distribución de la mezcla X se obtiene utilizando la fórmula de probabilidad total:

$$\begin{aligned} F_X(x) &= \mathbb{P}(X_M \le x) = \sum_{m\in\mathcal{M}} \mathbb{P}(X_M \le x | M = m)\mathbb{P}(M = m) \\ &= \sum_{m\in\mathcal{M}} \mathbb{P}(X_m \le x | M = m) p_M(m) \\ &= \sum_{m\in\mathcal{M}} \mathbb{P}(X_m \le x) p_M(m) \qquad \text{(pues } (X_m : m \in \mathcal{M}) \text{ y } M \text{ son indep.)} \\ &= \sum_{m\in\mathcal{M}} F_{X_m}(x) p_M(m), \end{aligned} \tag{5}$$

donde, para cada $m \in \mathcal{M}$, $F_{X_m}(x) = \mathbb{P}(X_m \le x)$ es la función de distribución de la variable X_m.

Variables discretas. Si las variables aleatorias X_m son discretas con funciones de probabilidad $p_{X_m}(x) = \mathbb{P}(X_m = x)$, respectivamente, la mezcla X es discreta y su función de probabilidad es

$$p_X(x) = \sum_{m \in \mathcal{M}} p_{X_m}(x) p_M(m). \tag{6}$$

Variables absolutamente continuas Si las variables X_m son absolutamente continuas con densidades $f_{X_m}(x)$, respectivamente, la mezcla X es absolutamente continua y tiene densidad

$$f_X(x) = \sum_{m \in \mathcal{M}} f_{X_m}(x) p_M(m). \tag{7}$$

Ejemplo 1.3. Para simular los valores de una variable aleatoria X se recurre al siguiente algoritmo: Se simula el valor de un variable aleatoria M con distribución Bernoulli de parámetro $p = 1/5$. Si $M = 0$, se simula el valor de una variable aleatoria X_0 con distribución uniforme sobre el intervalo $(0, 4)$. Si $M = 1$, se simula el valor de una variable aleatoria X_1 con distribución uniforme sobre el intervalo $(2, 6)$. Se quiere hallar la densidad de probabilidades de la variable X así simulada.

La variable X es una mezcla. La variable mezcladora es M y las variables aleatorias que componen la mezcla son X_0 y X_11. Por hipótesis, la variable mezcladora M se distribuye de acuerdo con la función de probabilidad

$$p_M(0) = 4/5, \qquad p_M(1) = 1/5$$

y las distribuciones de las variables componentes son $X_0 \sim \mathcal{U}(0, 4)$ y $X_1 \sim \mathcal{U}(2, 6)$. En otras palabras, las densidades de las variables componente son

$$f_{X_0}(x) = \frac{1}{4}\mathbf{1}\{0 < x < 4\} \qquad \text{y} \qquad f_{X_1}(x) = \frac{1}{4}\mathbf{1}\{2 < x < 6\}.$$

Usando la fórmula de probabilidad total (7) se obtiene la densidad de la mezcla X

$$\begin{aligned} f_X(x) &= p_M(0) f_{X_0}(x) + p_M(1) f_{X_1}(x) \\ &= \left(\frac{4}{5}\right)\frac{1}{4}\mathbf{1}\{0 < x < 4\} + \left(\frac{1}{5}\right)\frac{1}{4}\mathbf{1}\{2 < x < 6\} \\ &= \frac{4}{20}\mathbf{1}\{0 < x \leq 2\} + \frac{5}{20}\mathbf{1}\{2 < x < 4\} + \frac{1}{20}\mathbf{1}\{4 \leq x < 6\}. \end{aligned} \tag{8}$$

□

1.3. Sobre la regla de Bayes

Sean $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad; $M : \Omega \to \mathbb{R}$ una variable aleatoria discreta tal que $M(\Omega) = \mathcal{M}$ y $p_M(m) = \mathbb{P}(M = m) > 0$ para todo $m \in \mathcal{M}$ y $(X_m : m \in \mathcal{M})$ una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ e independiente de M. Supongamos además que las variables $X_m, m \in \mathcal{M}$ son absolutamente continuas con densidades de probabilidad continuas $f_{X_m}(x), m \in \mathcal{M}$, respectivamente.

Sea $X := X_M$ la mezcla de las variables M_m obtenida mediante la variable mezcladora M. ¿Qué sentido debería tener la expresión $\mathbb{P}(M = m|X = x)$? No debe olvidarse que la variable X es absolutamente continua y en consecuencia $\mathbb{P}(X = x) = 0$. Por lo tanto, no tiene ningún sentido definir $\mathbb{P}(M = m|X = x)$ mediante un cociente de la forma $\mathbb{P}(M = m|X = x) = \frac{\mathbb{P}(X=x, M=m)}{\mathbb{P}(X=x)} = \frac{0}{0}$.

¿Qué hacer? El obstáculo se puede superar siempre y cuando $f_X(x) > 0$. En tal caso, si "engordamos" el punto x mediante el intervalo de radio $h > 0$ (suficientemente chico) centrado en x, $B_h(x) := \{x - h < t < x + h\}$, el evento $\{X \in B_h(x)\}$ tiene probabilidad positiva

$$\mathbb{P}(X \in B_h(x)) = \int_{x-h}^{x+h} f_Y(t)dt = 2hf_X(\theta(h)), \qquad \theta(h) \in B_h(x). \tag{9}$$

y la probabilidad condicional del evento $\{M = m\}$, dado que ocurrió el evento $\{X \in B_h(x)\}$ está bien definida y vale

$$\mathbb{P}(M = m|X \in B_h(x)) = \frac{\mathbb{P}(M = m, X \in B_h(x))}{\mathbb{P}(X \in B_h(x))}.$$

Por otra parte,

$$\begin{aligned} \mathbb{P}(M = m,\, X \in B_h(x)) &= p_M(m)\mathbb{P}(X_m \in B_h(x)|M = m) \\ &= p_M(m)\mathbb{P}(X_m \in B_h(x)) \\ &= p_M(m) \int_{x-h}^{x+h} f_{X_m}(t)dt \\ &= 2hp_M(m)f_{X_m}(\theta_m(h)), \qquad \theta_m(h) \in B_h(x). \end{aligned} \tag{10}$$

De (9) y (10) se deduce que

$$\mathbb{P}\left(M = m|X \in B_h(x)\right) = \frac{p_M(m)f_{X_m}(\theta_m(h))}{f_X(\theta(h))} \tag{11}$$

Para "adelgazar" el punto "engordado" hacemos $h \to 0$ y obtenemos

$$\lim_{h\to 0} \mathbb{P}\left(M = m|X \in B_h(x)\right) = \lim_{h\to 0} \frac{p_M(m)f_{X_m}(\theta_m(h))}{f_X(\theta(h))} = \frac{p_M(m)f_{X_m}(x)}{f_X(x)}. \tag{12}$$

Finalmente, para cada $x \in \mathbb{R}$ tal que $f_X(x) > 0$ definimos $\mathbb{P}(M = m|X = x)$ mediante la fórmula

$$\mathbb{P}(M = m|X = x) := \frac{p_M(m)f_{X_m}(x)}{f_X(x)}. \tag{13}$$

Ejemplo 1.4 (Detección de señales)**.** Un emisor transmite un mensaje binario en la forma de una señal aleatoria Y que puede ser -1 o $+1$ con igual probabilidad. El canal de comunicación corrompe la transmisión con un ruido normal aditivo de media $\mu = 0$ y varianza 1. El receptor recibe la señal $X = N + Y$, donde N es un ruido (*noise*) con distribución $\mathcal{N}(0, 1)$, independiente de Y. La pregunta del receptor es la siguiente: dado que recibí el valor x, cuál es la probabilidad de que la señal sea 1?

La señal que recibe el receptor es una mezcla. La variable mezcladora es Y y las variables aleatorias que componen la mezcla son $X_{-1} = N-1$ y $X_1 = N+1$. Por hipótesis, la variable mezcladora S se distribuye de acuerdo con la función de probabilidad $p_Y(-1) = p_Y(1) = 1/2$ y las distribuciones de las variables componentes son $X_{-1} \sim \mathcal{N}(-1, 1)$ y $X_1 \sim \mathcal{N}(1, 1)$. En otras palabras, las densidades de las variables componente son

$$f_{X_{-1}}(x) = \frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2} \qquad \text{y} \qquad f_{X_1}(x) = \frac{1}{\sqrt{2\pi}} e^{-(z-1)^2/2}.$$

Usando la fórmula de probabilidad total (7) se obtiene la densidad de la mezcla X

$$f_X(x) \;=\; p_Y(-1) f_{X_{-1}}(x) + p_Y(1) f_{X_1}(x) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2}\right) + \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}} e^{-(z-1)^2/2}\right).$$

El receptor pregunta $\mathbb{P}(Y = 1|X = x) =?$ La respuesta se obtiene mediante la regla de Bayes (13)

$$\mathbb{P}(Y = 1|X = x) = \frac{p_Y(1) f_{X_1}(x)}{f_X(x)} = \frac{e^{-(x-1)^2/2}}{e^{-(x-1)^2/2} + e^{-(x+1)^2/2}} = \frac{e^x}{e^x + e^{-x}}. \tag{14}$$

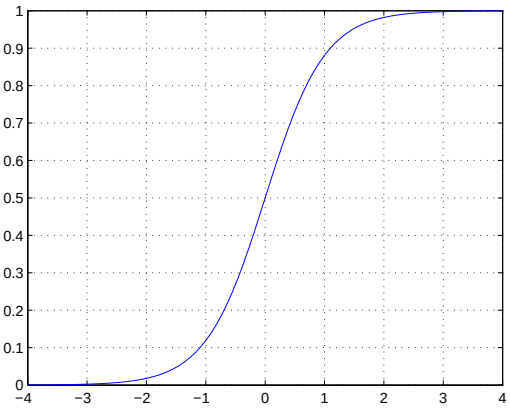


Figura 1: Gráfico de la probabilidad condicional $\mathbb{P}(Y = 1|X = \cdot) : \mathbb{R} \to \mathbb{R}$ vista como función de x.

□

1.4. Caso continuo

Sean X e Y dos variables aleatorias definidas sobre $(\Omega, \mathcal{A}, \mathbb{P})$ con densidad conjunta $f_{X,Y}(x,y)$ continua. A diferencia del caso en que X es discreta en este caso tenemos que $\mathbb{P}(X = x) = 0$ para todo $x \in \mathbb{R}$, lo que hace imposible definir la función de distribución condicional de Y dado que $X = x$, $\mathbb{P}(Y \leq y | X = x)$, mediante el cociente (2):

$$\frac{\mathbb{P}(Y \leq y, X = x)}{\mathbb{P}(X = x)} = \frac{0}{0}.$$

Este obstáculo se puede superar observando que para cada $x \in Sop(f_X)$ y para cada $h > 0$ el evento $\{X \in B_h(x)\} = \{x - h < X < x + h\}$ tiene probabilidad positiva

$$\mathbb{P}(X \in B_h(x)) = \int_{x-h}^{x+h} f_X(s)ds = 2h f_X(\theta_1(h)), \qquad \theta_1(h) \in B_h(x).$$

Por otra parte,

$$\mathbb{P}(Y \leq y, X \in B_h(x)) = \int_{x-h}^{x+h} \left(\int_{-\infty}^{y} f_{X,Y}(s,t)dt \right) ds = 2h \int_{-\infty}^{y} f_{X,Y}(\theta_2(h), t)dt,$$

donde $\theta_2(h) \in B_h(x)$.

Si $x \in Sop(f_X)$, la probabilidad condicional $\mathbb{P}(Y \leq y | X \in B_h(x))$ está bien definida y vale

$$\mathbb{P}\left(Y \leq y | X \in B_h(x)\right) = \frac{\mathbb{P}(Y \leq y,\ X \in B_h(x))}{\mathbb{P}(X \in B_h(x))} = \frac{\int_{-\infty}^{y} f_{X,Y}(\theta_2(h), t)dt}{f_X(\theta_1(h))}.$$

En consecuencia,

$$\lim_{h \to 0} \mathbb{P}(Y \leq y | X \in B_h(x)) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,t)dt}{f_X(x)}. \tag{15}$$

El lado derecho de (15) define una genuina función de distribución $F_{Y|X}(\cdot\,|x) : \mathbb{R} \to \mathbb{R}$,

$$F_{Y|X}(y\,|x) := \frac{\int_{-\infty}^{y} f_{X,Y}(x,t)dt}{f_X(x)}, \tag{16}$$

que se llama la *función distribución condicional de Y dado $X = x$* y se puede interpretar como la función de distribución de una nueva variable aleatoria que llamaremos Y *condicional a que $X = x$* y que será designada mediante el símbolo $Y|X = x$.

La función de distribución $F_{Y|X}(y|x)$ es derivable y su derivada

$$f_{Y|X}(y|x) := \frac{d}{dy} F_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{17}$$

se llama la *densidad condicional de Y dado que $X = x$*.

Curva peligrosa. Todo el argumento usa la hipótesis $f_X(x) > 0$. Si $f_X(x) = 0$ las expresiones (15)-(17) carecen de sentido. Sin embargo, esto no es un problema grave ya que $\mathbb{P}(X \in Sop(f_X)) = 1$. Para los valores de x tales que $f_X(x) = 0$ las variables condicionales $Y|X = x$ serán definidas como idénticamente nulas. En tal caso, $F_{Y|X}(y|x) = \mathbf{1}\{y \geq 0\}$ y $f_{Y|X}(y|x) = \delta_0(y)$.

Regla mnemotécnica. De la fórmula (17) se deduce que $f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x)$ y puede recordarse mediante el siguiente "versito": "*la densidad conjunta es igual a la densidad condicional por la marginal de la condición*".

Ejemplo 1.5 (Dos etapas: conjunta = marginal × condicional)**.** Se elige un número al azar, X en el intervalo $(0, 1)$ y después otro número al azar, Y, en el intervalo $(X, 1)$. Se quiere hallar la densidad marginal de Y. Por hipótesis, $X \sim \mathcal{U}(0, 1)$ e $Y \sim \mathcal{U}(X, 1)$:

$$f_X(x) = \mathbf{1}\{0 < x < 1\} \qquad \text{y} \qquad f_{Y|X}(y|x) = \frac{1}{1-x}\mathbf{1}\{x < y < 1\}.$$

La densidad conjunta de X e Y se obtiene multiplicando la densidad condicional $f_{Y|X}(y|x)$ por la densidad marginal $f_X(x)$

$$f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x) = \frac{1}{1-x}\mathbf{1}\{0 < x < y < 1\}.$$

La densidad marginal de Y se obtiene integrando la densidad conjunta $f_{X,Y}(x, y)$ con respecto a x

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} \frac{1}{1-x}\mathbf{1}\{0 < x < y < 1\}dx = \mathbf{1}\{0 < y < 1\} \int_0^y \frac{1}{1-x}dx \\ &= -\log(1-y)\mathbf{1}\{0 < y < 1\}. \end{aligned}$$

□

Fórmula de probabilidad total. La densidad de probabilidades de Y es una combinación convexa de las condicionales:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx.$$

Inmediato de la relación "conjunta = marginal × condicional":

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx.$$

Integrando respecto de y se obtiene que la función de distribución de Y es una combinación convexa de las condicionales:

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y} \left(\int_{-\infty}^{\infty} f_{Y|X}(t|x)f_X(x)dx \right) dt \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f_{Y|X}(t|x)dt \right) f_X(x)dx = \int_{-\infty}^{\infty} F_{Y|X}(y|x)f_X(x)dx. \end{aligned}$$

Esperanza condicional de Y dado que $X = x$. Para cada $x \in \mathbb{R}$, la *esperanza condicional de Y dado que $X = x$* se define por

$$\mathbb{E}[Y|X = x] := \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy. \tag{18}$$

siempre y cuando la integral del converja absolutamente. Observar que si $f_X(x) = 0$, $\mathbb{E}[X|X = x] = 0$.

Línea de regresión. Considerando a x como una *variable*, el lado derecho de (18) define una función $\varphi : \mathbb{R} \to \mathbb{R}$

$$\varphi(x) := \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

cuyo gráfico se denomina la *línea de regresión de Y sobre X*, abusando del lenguaje la llamaremos del mismo modo.

Varianza condicional

En cualquier caso, definidas las esperanzas condicionales de Y y de Y^2 dado que $X = x$, la *varianza condicional de Y dado que $X = x$* se define mediante

$$\mathbb{V}(Y|X = x) := \mathbb{E}\left[(Y - \mathbb{E}[Y|X = x])^2 \,|X = x\right] \tag{19}$$

Desarrollando el término derecho se obtiene

$$\mathbb{V}(Y|X = x) = \mathbb{E}[Y^2|X = x] - \mathbb{E}[Y|X = x]^2. \tag{20}$$

Nota Bene. La definición es consistente y coincide con la varianza de la variable aleatoria $Y|X = x$ cuya función de distribución es $F_{Y|X}(y|x)$.

Ejemplo 1.6 (Dardos)**.** Volvamos al problema del juego de dardos de blanco circular $\Lambda = \{(x, y) \in \mathbb{R}^2 : x^2+y^2 \leq 1\}$. Por hipótesis, el dardo se clava en un punto de coordenadas (X, Y) uniformemente distribuido sobre Λ. Esto significa que la densidad conjunta de X e Y es

$$f_{X,Y}(x, y) = \frac{1}{\pi}\mathbf{1}\{x^2 + y^2 \leq 1\}.$$

Por definición, para cada $x \in [-1, 1]$, la densidad condicional de Y dado que $X = x$ es el cociente entre la densidad conjunta $f_{X,Y}(x, y)$ y la densidad marginal de X

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

Por lo tanto,

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{1 - x^2}}\mathbf{1}\{-\sqrt{1 - x^2} \leq y \leq \sqrt{1 - x^2}\}. \tag{21}$$

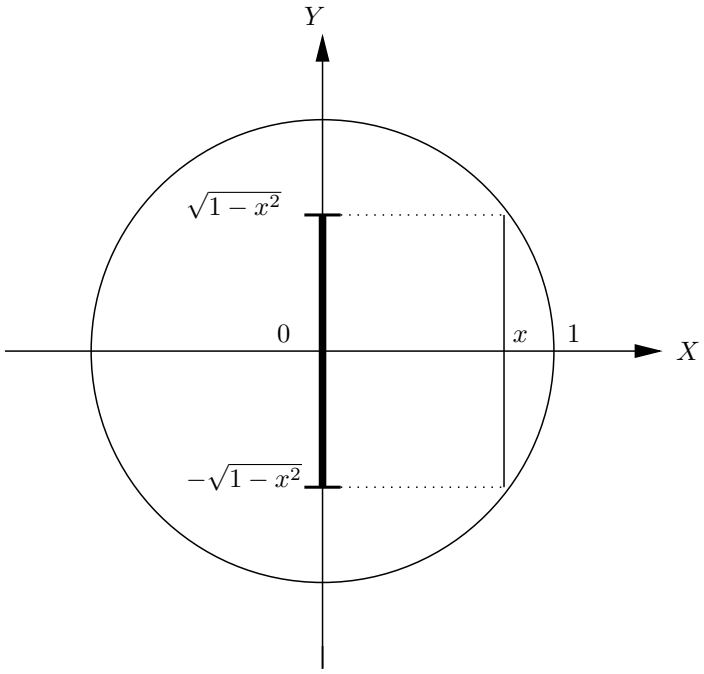


Figura 2: Para cada $x \in [-1,1]$ se observa que $Y|X=x \sim \mathcal{U}\left[-\sqrt{1-x^2}, \sqrt{1-x^2}\right]$.

En otras palabras, dado que $X = x$, $x \in [-1,1]$, la variable Y se distribuye uniformemente sobre el intervalo $\left[-\sqrt{1-x^2}, \sqrt{1-x^2}\right]$. En consecuencia,

$$\mathbb{E}[Y|X=x] = 0 \qquad \text{y} \qquad \mathbb{V}(Y|X=x) = (2\sqrt{1-x^2})^2/12 = (1-x^2)/3.$$

□

Ejemplo 1.7. Observando el Cuadro 2 correspondiente al Ejemplo 1.1 se deduce que

$$\mathbb{E}[Y^2|X=x] = \frac{6}{3}\mathbf{1}\{x=0\} + \frac{2}{3}\mathbf{1}\{x=1\}.$$

De (20) y la expresión de $\mathbb{E}[Y|X=x]$ obtenida en (4) resulta que

$$\begin{aligned}\mathbb{V}(Y|X=x) &= \frac{6}{3}\mathbf{1}\{x=0\} + \frac{2}{3}\mathbf{1}\{x=1\} - \left(\frac{4}{3}\mathbf{1}\{x=0\} + \frac{2}{3}\mathbf{1}\{x=1\}\right)^2 \\ &= \frac{2}{9}\mathbf{1}\{x=0\} + \frac{5}{9}\mathbf{1}\{x=1\}.\end{aligned}$$

□

2. Predicción y Esperanza condicional

Planteo del problema

En su versión más simple un problema de predicción o estimación involucra dos variables aleatorias: una variable aleatoria Y desconocida (o inobservable) y una variable aleatoria X conocida (u observable). El problema consiste en deducir información sobre el valor de Y a partir del conocimiento del valor de X. Para ser más precisos, se busca una función $\varphi(X)$ que (en algún sentido) sea lo más parecida a Y como sea posible. La variable aleatoria $\hat{Y} := \varphi(X)$ se denomina un *estimador* de Y.

Ejemplo 2.1 (Detección de señales)**.** Un emisor transmite un mensaje binario en la forma de una señal aleatoria Y que puede ser -1 o $+1$ con igual probabilidad. El canal de comunicación corrompe la transmisión con un ruido normal aditivo de media $\mu = 0$ y varianza σ^2. El receptor recibe la señal $X = Y + N$, donde N es un ruido (*noise*) con distribución $\mathcal{N}(0, \sigma^2)$, independiente de Y. El receptor del mensaje observa la señal corrompida X y sobre esa base tiene que "reconstruir" la señal original Y. ¿Cómo lo hace?, ¿Qué puede hacer?

En lo que sigue desarrollaremos herramientas que permitan resolver este tipo de problemas. Sean X e Y dos variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. El objetivo es construir una función $\varphi(X)$ que *sea lo más parecida a Y como sea posible*. En primer lugar, vamos a suponer que $\mathbb{E}[|Y|] < \infty$. Esta hipótesis permite precisar el sentido del enunciado *parecerse a Y*. Concretamente, queremos construir una función de X, $\varphi(X)$, que solucione la siguiente ecuación funcional

$$\mathbb{E}[\varphi(X)h(X)] = \mathbb{E}[Yh(X)], \tag{22}$$

para toda función medible y acotada $h : \mathbb{R} \to \mathbb{R}$.

Esperanza condicional

Sean X e Y dos variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Supongamos que $\mathbb{E}[|Y|] < \infty$. Definimos la *esperanza condicional de Y dada X*, $\mathbb{E}[Y|X]$, como cualquier variable aleatoria de la forma $\varphi(X)$, donde $\varphi : \mathbb{R} \to \mathbb{R}$ es una función (medible), que solucione la ecuación funcional (22).

Existencia. La existencia de la esperanza condicional depende de teoremas profundos de Teoría de la medida y no será discutida en estas notas. El lector interesado puede consultar Billingsley(1986) y/o Durrett(1996).

Unicidad. Supongamos que $\varphi(X)$ y $\psi(X)$ son dos soluciones de la ecuación funcional (22). Entonces, $\varphi(X) = \psi(X)$ casí seguramente (i.e., $\mathbb{P}(\varphi(X) \neq \psi(X)) = 0$).

Demostración. Por cuestiones de simetría, la prueba se reduce a mostrar que para cada $\epsilon > 0$, $\mathbb{P}(A_\epsilon) = 0$, donde $A_\epsilon := \{\varphi(X) - \psi(X) \geq \epsilon\}$. Observar que, por hipótesis, para toda función medible y acotada $h : \mathbb{R} \to \mathbb{R}$ vale que

$$\mathbb{E}[\varphi(X)h(X)] = \mathbb{E}[\psi(X)h(X)] \iff \mathbb{E}[(\varphi(X) - \psi(X))h(X)] = 0.$$

Poniendo $h(X) = \mathbf{1}\{X \in A_\epsilon\}$ se observa que

$$0 = \mathbb{E}[(\varphi(X) - \psi(X))\mathbf{1}\{X \in A_\epsilon\}] \geq \mathbb{E}[\epsilon\mathbf{1}\{X \in A_\epsilon\}] = \epsilon\mathbb{P}(A_\epsilon).$$

Por lo tanto, $\mathbb{P}(A_\epsilon) = 0$. □

Lema 2.2 (Técnico)**.** La esperanza condicional satisface $\mathbb{E}[|\mathbb{E}[Y|X]|] \leq \mathbb{E}[|Y|]$.

Demostración. La variable aleatoria $\varphi(X)$ es solución de la ecuación funcional (22). El conjunto $\{\varphi(X) > 0\} \in \mathcal{A}$. Poniendo $h(X) = \mathbf{1}\{\varphi(X) > 0\}$ y usando (22) se obtiene

$$\mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) > 0\}] = \mathbb{E}[Y\mathbf{1}\{\varphi(X) > 0\}] \leq \mathbb{E}[|Y|].$$

Análogamente puede verse que

$$\mathbb{E}[-\varphi(X)\mathbf{1}\{\varphi(X) \leq 0\}] = \mathbb{E}[-Y\mathbf{1}\{\varphi(X) \leq 0\}] \leq \mathbb{E}[|Y|]$$

Por lo tanto,

$$\begin{aligned} E[|\varphi(X)|] &= \mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) > 0\} - \varphi(X)\mathbf{1}\{\varphi(X) \leq 0\}] \\ &= \mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) > 0\}] - \mathbb{E}[\varphi(X)\mathbf{1}\{\varphi(X) \leq 0\}] \\ &= \mathbb{E}[Y\mathbf{1}\{\varphi(X) > 0\}] - \mathbb{E}[Y\mathbf{1}\{\varphi(X) \leq 0\}] \\ &= \mathbb{E}[Y\mathbf{1}\{\varphi(X) > 0\} - Y\mathbf{1}\{\varphi(X) \leq 0\}] \\ &\leq \mathbb{E}[|Y|]]. \end{aligned}$$

□

Propiedades que merecen ser subrayadas

Aunque se deducen inmediatamente de la definición, las propiedades siguientes merecen ser subrayas porque, como se podrá apreciar más adelante, constituyen poderosas herramientas de cálculo.

1. Fórmula de probabilidad total:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]. \tag{23}$$

2. Sea $g : \mathbb{R} \to \mathbb{R}$ una función tal que $\mathbb{E}[|g(X)Y|] < \infty$,

$$\mathbb{E}[g(X)Y|X] = g(X)\mathbb{E}[Y|X]. \tag{24}$$

3. Si X e Y son independientes, entonces $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.

Demostración. La fórmula de probabilidad total se deduce de la ecuación (22) poniendo $h(X) \equiv 1$. La identidad (24) se obtiene observando que $g(X)\mathbb{E}[Y|X]$ es una función de X que soluciona la ecuación $\mathbb{E}[g(X)\mathbb{E}[Y|X]h(X)] = \mathbb{E}[(g(X)Y)h(X)]$. Si X e Y son independientes $\mathbb{E}[Yh(X)] = \mathbb{E}[Y]\mathbb{E}[h(X)] = \mathbb{E}[\mathbb{E}[Y]h(X)]$. □

2.1. Ejemplos

2.1.1. Caso continuo

Sean X e Y dos variables aleatorias continuas definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, con densidad de probabilidades conjunta $f_{X,Y}(x, y)$ y $\mathbb{E}[|Y|] < \infty$. La esperanza condicional de Y dada X es $\mathbb{E}[Y|X] = \varphi(X)$, donde $\varphi : \mathbb{R} \to \mathbb{R}$ es la función de regresión de Y sobre X definida por

$$\varphi(x) := \mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy. \tag{25}$$

Demostración. Basta ver $\varphi(X)$ verifica la ecuación funcional (22) para cualquier función h medible y acotada.

$$\begin{aligned}
\mathbb{E}[\varphi(X)h(X)] &= \int_{-\infty}^{\infty} \varphi(x)h(x)f_X(x)dx \\
&= \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x]h(x)f_X(x)dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \right) h(x)f_X(x)dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yh(x)f_{Y|X}(y|x)f_X(x)dxdy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yh(x)f_{X,Y}(x, y)dxdy \\
&= \mathbb{E}[Yh(X)].
\end{aligned}$$

□

2.1.2. Regla de Bayes para mezclas

Volvamos el Ejemplo 2.1 la pregunta es ¿Qué puede hacer el receptor para "reconstruir" la señal original, Y, a partir de la señal corrompida X? Lo "mejor" que puede hacer es estimar Y mediante la esperanza condicional $\mathbb{E}[Y|X]$.

Recordemos brevemente las condiciones del problema. El emisor transmite un mensaje binario en la forma de una señal aleatoria Y que puede ser -1 o $+1$ con igual probabilidad. El canal de comunicación corrompe la transmisión con un ruido normal aditivo de media $\mu = 0$ y varianza σ^2. El receptor recibe la señal $X = Y + N$, donde N es un ruido con distribución $\mathcal{N}(0, \sigma^2)$, independiente de Y.

El receptor recibe la mezcla de dos variables aleatorias $X|Y=-1 \sim \mathcal{N}(-1,\sigma^2)$ e $Y|X=1 \sim \mathcal{N}(1,\sigma^2)$, mezcladas en igual proporción: $p_X(-1)=p_X(1)=1/2$. Las densidades de las componentes de la mezcla son

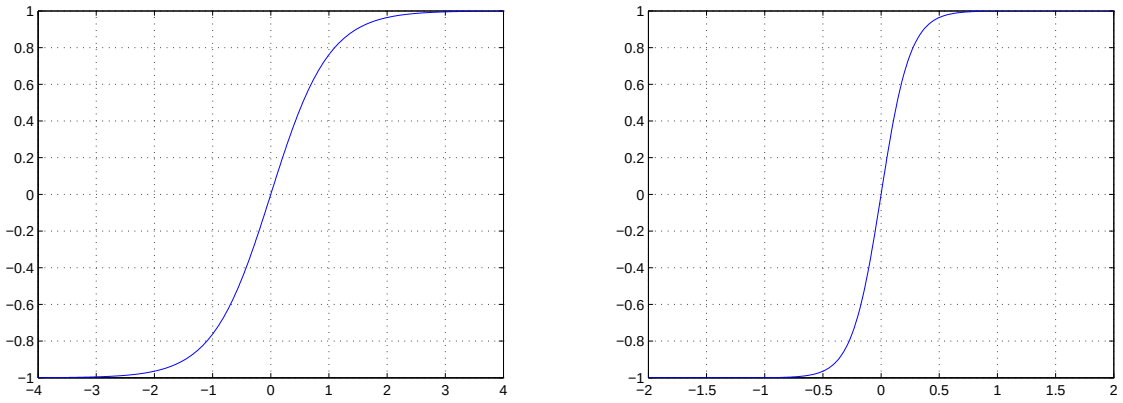
$$f_{X|Y}(x|-1)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(x+1)^2/2\sigma^2} \qquad \text{y} \qquad f_{X|Y}(x|1)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(x-1)^2/2\sigma^2}.$$

De la fórmula de probabilidad total se deduce que la densidad de la mezcla X es

$$\begin{aligned} f_X(x) &= p_Y(-1)f_{X|Y}(x|-1)+p_Y(1)f_{X|Y}(x|1) \\ &= \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(x+1)^2/2\sigma^2}\right)+\frac{1}{2}\left(\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(x-1)^2/2\sigma^2}\right). \end{aligned} \tag{26}$$

Para construir la esperanza condicional $\mathbb{E}[Y|X]$ el receptor debe calcular la línea de regresión $\varphi(x)=\mathbb{E}[Y|X=x]=1\mathbb{P}(Y=1|X=x)-1\mathbb{P}(Y=-1|X=x)$. Que de acuerdo con la regla de Bayes para mezclas adopta la forma

$$\varphi(x)=\frac{p_Y(1)f_{X|Y}(x|1)-p_Y(-1)f_{X|Y}(x|-1)}{f_X(x)}=\frac{e^{x/\sigma^2}-e^{-x\sigma^2}}{e^{x/\sigma^2}+e^{-x\sigma^2}}=\tanh(x/\sigma^2) \tag{27}$$



(a) (b)

Figura 3: Líneas de regresión de Y sobre X para distintos valores de la varianza σ^2. (a) $\sigma^2=1$: $\varphi(x)=\tanh(x)$; (b) $\sigma^2=1/4$, $\varphi(x)=\tanh(4x)$.

Finalmente, el receptor reconstruye Y basándose en X mediante $\mathbb{E}[Y|X]=\tanh(X/\sigma^2)$. □

2.1.3. Caso discreto

Sean X e Y dos variables aleatorias discretas definidas sobre un mismo espacio de probabilidad $(\Omega,\mathcal{A},\mathbb{P})$,con función de probabilidad conjunta $p_{X,Y}(x,y)$ y $\mathbb{E}[|Y|]<\infty$. Para

simplificar la exposición supongamos que $Sop(p_X) = X(\Omega)$. En tal caso, la esperanza condicional de Y dada X es $\mathbb{E}[Y|X] = \varphi(X)$, donde $\varphi : \mathbb{R} \to \mathbb{R}$ es la línea de regresión de Y sobre X definida por

$$\varphi(x) := \mathbb{E}[Y|X = x] = \sum_{y \in Y(\Omega)} y p_{Y|X}(y|x) \tag{28}$$

Demostración. Basta ver $\varphi(X)$ verifica la ecuación funcional (22) para cualquier función h medible y acotada.

$$\begin{aligned}
\mathbb{E}[\varphi(X)h(X)] &= \sum_x \varphi(x)h(x)p_X(x) = \sum_x \mathbb{E}[Y|X = x]h(x)p_X(x) \\
&= \sum_x \left(\sum_y y p_{Y|X}(y|x) \right) h(x)p_X(x) = \sum_x \sum_y y h(x) p_{Y|X}(y|x)p_X(x) \\
&= \sum_x \sum_y y h(x) p_{X,Y}(x, y) = \mathbb{E}[Y h(X)].
\end{aligned}$$

□

Ejemplo 2.3 (Fórmula de probabilidad total)**.** Una rata está atrapada en un laberinto. Inicialmente puede elegir una de tres direcciones. Si elige la primera se perderá en el laberinto y luego de 4 minutos volverá a su posición inicial; si elige la segunda volverá a su posición inicial luego de 7 minutos; si elige la tercera saldrá del laberinto luego de 3 minutos. Suponiendo que en cada intento, la rata elige con igual probabilidad cualquiera de las tres direcciones, cuál es la esperanza del tiempo que demora en salir del laberinto?

Sean Y la cantidad de tiempo que demora la rata en salir del laberinto y sea X la dirección que elige inicialmente. Usando la fórmula de probabilidad total puede verse que

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \sum_{x=1}^{3} \mathbb{E}[Y|X = x]\mathbb{P}(X = x) = \frac{1}{3} \sum_{x=1}^{3} \mathbb{E}[Y|X = x]$$

Si la rata elige la primera dirección, se pierde en el laberinto durante 4 minutos y vuelve a su posición inicial. Una vez que vuelve a su posición inicial el problema se renueva y la esperanza del tiempo adicional hasta que la rata consiga salir del laberinto es $\mathbb{E}[Y]$. En otros términos $\mathbb{E}[Y|X = 1] = 4 + \mathbb{E}[Y]$. Análogamente puede verse que $\mathbb{E}[Y|X = 2] = 7 + \mathbb{E}[Y]$. La igualdad $\mathbb{E}[Y|X = 3] = 3$ no requiere comentarios. Por lo tanto,

$$\mathbb{E}[Y] = \frac{1}{3}\left(4 + \mathbb{E}[Y] + 7 + \mathbb{E}[Y] + 3\right) = \frac{1}{3}\left(2\mathbb{E}[Y] + 14\right)$$

Finalmente, $\mathbb{E}[Y] = 14$. □

2.2. Propiedades

La esperanza condicional tiene propiedades similares a la esperanza.

Linealidad. $\mathbb{E}[aY_1 + bY_2|X] = a\mathbb{E}[Y_1|X] + b\mathbb{E}[Y_2|X]$.

Monotonía. Si $Y_1 \leq Y_2$, entonces $\mathbb{E}[Y_1|X] \leq \mathbb{E}[Y_2|X]$.

Desigualdad de Jensen. Si $g : \mathbb{R} \to \mathbb{R}$ es una función convexa y $\mathbb{E}[|Y|], \mathbb{E}[|g(Y)|] < \infty$, entonces

$$g(\mathbb{E}[Y|X]) \leq \mathbb{E}[g(Y)|X]. \tag{29}$$

En particular, si $\mathbb{E}[Y^2] < \infty$, poniendo $g(t) = t^2$ en la desigualdad de Jensen se obtiene

$$\mathbb{E}[Y|X]^2 \leq \mathbb{E}[Y^2|X] \tag{30}$$

Definición 2.4 (Varianza condicional)**.** Sean X e Y dos variables aleatorias sobre el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$. Si $\mathbb{E}[Y^2] < \infty$, la *varianza condicional de Y dada X*, $\mathbb{V}(Y|X)$, se define por

$$\mathbb{V}(Y|X) = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2 \tag{31}$$

Predicción

Existen diversas maneras en las que dos variables pueden considerarse cercanas entre sí. Una manera es trabajar con la norma dada por $\|X\| := \sqrt{\mathbb{E}[X^2]}$ y definir la distancia entre dos variables aleatorias X e Y, $d(X, Y)$ mediante

$$d(X, Y) := \|Y - X\| = \sqrt{\mathbb{E}[(Y - X)^2]}. \tag{32}$$

.

Definición 2.5 (Predictor)**.** Sean X e Y variables aleatorias sobre un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, tales que $\mathbb{E}[Y^2] < \infty$. El predictor de error cuadrático medio mínimo (o *mejor predictor*) de Y dada X es la función $\hat{Y} = h(X)$ de X que minimiza la distancia $d(\hat{Y}, Y)$ definida en (32).

El mejor predictor de Y dada X es una variable aleatoria $\hat{Y}$ perteneciente al espacio vectorial $\mathbb{H} = \{h(X) : h : \mathbb{R} \to \mathbb{R}, \mathbb{E}[h(X)^2] < \infty\}$ tal que $\mathbb{E}[(Y - \hat{Y})^2] \leq \mathbb{E}[(Y - Z)^2]$ para toda $Z \in \mathbb{H}$.

Interpretación geométrica. Sea $L_2(\Omega, \mathcal{A}, \mathbb{P})$ el conjunto de todas la variables aleatorias sobre $(\Omega, \mathcal{A}, \mathbb{P})$ que tienen varianza finita. $\mathbb{H}$ es un subespacio de $L_2(\Omega, \mathcal{A}, \mathbb{P})$. Si $Y \notin \mathbb{H}$ entonces el camino más corto desde Y hasta $\mathbb{H}$ es por la recta ortogonal al subespacio $\mathbb{H}$ que pasa por Y. Por lo tanto $\hat{Y}$ debe ser la proyección ortogonal de Y sobre $\mathbb{H}$. En tal caso $Y - \hat{Y}$ es ortogonal a cualquier vector de $\mathbb{H}$. En otras palabras, $\langle Y - \hat{Y}, Z\rangle = 0$ para todo $Z \in \mathbb{H}$, donde $\langle X, Y\rangle$ es el producto interno en $L_2(\Omega, \mathcal{A}, \mathbb{P})$ definido por $\langle X, Y\rangle := \mathbb{E}[XY]$.

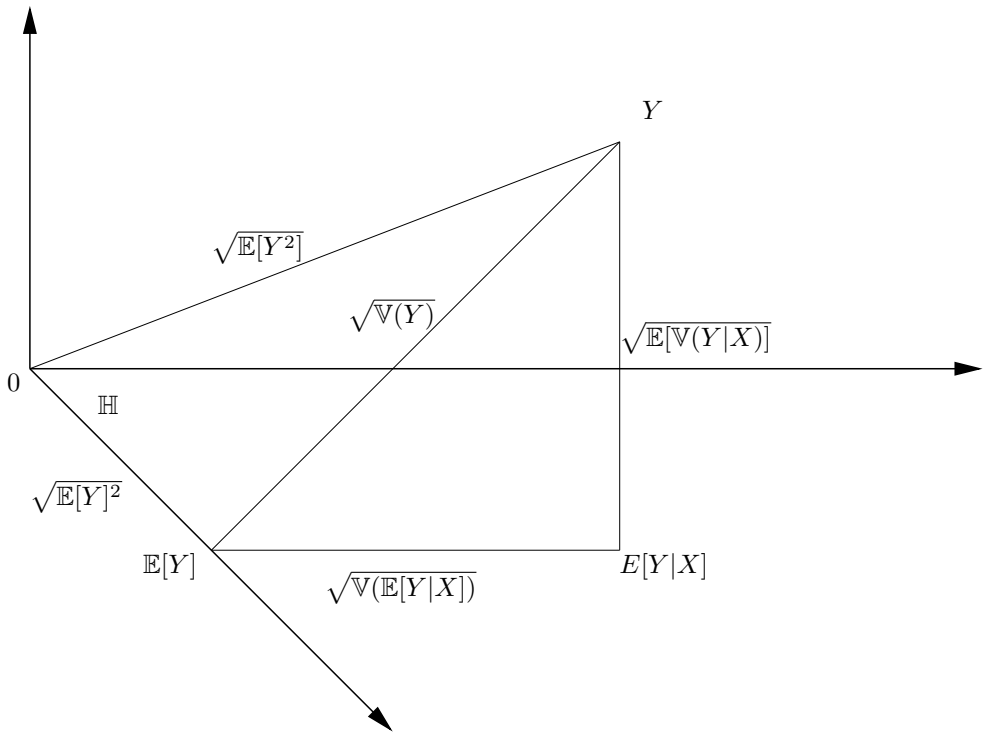


Figura 4: Teorema de Pitágoras: $\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}(\mathbb{E}[Y|X])$.

La esperanza condicional $\mathbb{E}[Y|X]$ es el mejor predictor de Y basado en X

1) La condición $\mathbb{E}[Y^2] < \infty$ implica que $\mathbb{E}[Y|X] \in \mathbb{H}$:

$$\mathbb{E}[\mathbb{E}[Y|X]^2] \leq \mathbb{E}[\mathbb{E}[Y^2|X]] = \mathbb{E}[Y^2] < \infty.$$

2) La ecuación funcional (22) significa que $Y - \mathbb{E}[Y|X] \perp \mathbb{H}$:

$$\begin{aligned} \langle Y - \mathbb{E}[Y|X],\, h(X)\rangle = 0 \quad &\Longleftrightarrow \quad \mathbb{E}[(Y - \mathbb{E}[Y|X])h(X)] = 0 \\ &\Longleftrightarrow \quad \mathbb{E}[\mathbb{E}[Y|X]h(X)] = \mathbb{E}[Yh(X)]. \end{aligned}$$

Por lo tanto, la esperanza condicional, $\mathbb{E}[Y|X]$, satisface las dos condiciones que caracterizan a la proyección ortogonal sobre el subespacio $\mathbb{H}$ y en consecuencia es el predictor de Y basado en X de error cuadrático mínimo:

$$\mathbb{E}[Y|X] = \arg \min_{h(X)\in\mathbb{H}} \mathbb{E}[(Y - h(X))^2].$$

El error cuadrático medio mínimo se puede expresar en la forma

$$\begin{aligned} \|Y - \mathbb{E}[Y|X]\|^2 &= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] \\ &= \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]] \\ &= \mathbb{E}[\mathbb{V}(Y|X)]. \end{aligned}$$

La última igualdad se obtiene desarrollando el cuadrado $(Y - \mathbb{E}[Y|X])^2$ y usando las propiedades de la esperanza condicional. (*Ejercicio*)

Por último, como $\mathbb{E}[Y] \in \mathbb{H}$, el Teorema de Pitágoras implica que

$$\begin{aligned} \mathbb{V}(Y) &= \|Y - \mathbb{E}[Y]\|^2 = \|Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - \mathbb{E}[Y]\|^2 \\ &= \|Y - \mathbb{E}[Y|X]\|^2 + \|\mathbb{E}[Y|X] - \mathbb{E}[Y]\|^2 \\ &= \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}(\mathbb{E}[Y|X]). \end{aligned} \tag{33}$$

En otras palabras, la variabilidad de Y se descompone de la siguiente manera: la variabilidad (media) de Y alrededor de su esperanza condicional, más la variabilidad de esta última.

2.3. Ejemplo: sumas aleatorias de variables aleatorias

Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias idénticamente distribuidas de media μ y varianza σ^2. Sea N una variable discreta a valores en $\mathbb{N}$ que es independiente de las X_i. El problema consiste en hallar la media y la varianza de la variable aleatoria $S = \sum_{i=1}^{N} X_i$, llamada *variable aleatoria compuesta*. Este problema se puede resolver utilizando las identidades

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|N]] \qquad \text{y} \qquad \mathbb{V}(S) = \mathbb{E}[\mathbb{V}(S|N)] + \mathbb{V}(\mathbb{E}[S|N]).$$

En la jerga probabilística esta técnica de cálculo se conoce bajo el nombre de *cálculo de esperanzas y varianzas mediante condicionales*.

Cálculo de la esperanza por condicionales.

$$\begin{aligned} \mathbb{E}\left[S|N=n\right] &= \mathbb{E}\left[\sum_{i=1}^{N} X_i \middle| N=n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i \middle| N=n\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \qquad \text{por la independencia de las } X_i \text{ y } N \\ &= n\mu. \end{aligned}$$

En consecuencia,

$$\mathbb{E}\left[S|N\right] = \mu N.$$

Por lo tanto,

$$\mathbb{E}\left[S\right] = \mathbb{E}[\mathbb{E}[S|N]] = \mathbb{E}\left[\mu N\right] = \mu\,\mathbb{E}[N].$$

□

Cálculo de la varianza por condicionales.

$$\begin{aligned}\mathbb{V}(S|N=n) &= \mathbb{V}\left(\sum_{i=1}^{N} X_i \big|\, N=n\right) = \mathbb{V}\left(\sum_{i=1}^{n} X_i \big|\, N=n\right)\\ &= \mathbb{V}\left(\sum_{i=1}^{n} X_i\right) \qquad \text{por la independencia de } X_i \text{ y } N\\ &= n\sigma^2.\end{aligned}$$

En consecuencia,

$$\mathbb{V}(S|N) = \sigma^2 N.$$

Por lo tanto,

$$\mathbb{E}[\mathbb{V}(S|N)] = \mathbb{E}[\sigma^2 N] = \sigma^2\,\mathbb{E}[N].$$

Por otra parte,

$$\mathbb{V}[\mathbb{E}(S|N)] = \mathbb{V}[\mu N] = \mu^2\,\mathbb{V}[N].$$

Finalmente,

$$\mathbb{V}(S) = \mathbb{E}[\mathbb{V}(S|N)] + \mathbb{V}(\mathbb{E}[S|N]) = \sigma^2\,\mathbb{E}[N] + \mu^2\,\mathbb{V}[N].$$

□

2.4. Ejemplo: esperanza y varianza de una mezcla.

Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad. Sea $M : \Omega \to \mathbb{R}$ una variable aleatoria discreta tal que $M(\Omega) = \mathcal{M}$ y $p_M(m) = \mathbb{P}(M = m) > 0$ para todo $m \in \mathcal{M}$ y sea $(X_m : m \in \mathcal{M})$ una familia de variables aleatorias definidas sobre el mismo espacio de probabilidad, independiente de M. El problema consiste en hallar la media y la varianza de la mezcla $X := X_M$.

La forma natural de resolver este problema es usar la técnica del *cálculo de esperanzas y varianzas mediante condicionales*:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|M]] \qquad \text{y} \qquad \mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|M)] + \mathbb{V}(\mathbb{E}[X|M]).$$

Cálculo de la esperanza por condicionales. En primer lugar hay que observar que $X|M = m \sim X_m$ por lo tanto,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|M]] = \sum_{m\in\mathcal{M}} \mathbb{E}\left[X|M=m\right]\mathbb{P}(M=m) = \sum_{m\in\mathcal{M}} \mathbb{E}[X_m]p_M(m).$$

□

Cálculo de la varianza por condicionales.

$$\mathbb{E}[\mathbb{V}(X|M)] = \sum_{m\in\mathcal{M}} \mathbb{V}(X|M=m)\mathbb{P}(M=m) = \sum_{m\in\mathcal{M}} \mathbb{V}(X_m)p_M(m).$$

Por otra parte,

$$\begin{aligned}\mathbb{V}(\mathbb{E}[X|M]) &= \mathbb{E}[(\mathbb{E}[X|M]-\mathbb{E}[X])^2] = \sum_{m\in\mathcal{M}} (\mathbb{E}[X|M=m]-\mathbb{E}[X])^2\mathbb{P}(M=m)\\ &= \sum_{m\in\mathcal{M}} (\mathbb{E}[X_m]-\mathbb{E}[X])^2 p_M(m).\end{aligned}$$

Finalmente,

$$\mathbb{V}(X) = \sum_{m\in\mathcal{M}} \mathbb{V}(X_m)p_M(m) + \sum_{m\in\mathcal{M}} (\mathbb{E}[X_m]-\mathbb{E}[X])^2 p_M(m).$$

Nota Bene. Comparar con el Teorema de Steiner para el momento de inercia. □

3. Predicción lineal y coeficiente de correlación

Definición 3.1 (Predictor lineal)**.** Sean X e Y variables aleatorias sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, tales que $\mathbb{E}[X^2] < \infty$ y $\mathbb{E}[Y^2] < \infty$. La *recta de regresión de Y basada en X* es la función lineal $\hat{Y} = aX + b$ que minimiza la distancia

$$d(\hat{Y}, Y) = \sqrt{\mathbb{E}[(Y-\hat{Y})^2]}.$$

Cálculo explícito de la recta de regresión. El problema consiste en hallar los valores de a y b que minimizan la siguiente función de dos variables

$$g(a,b) := \mathbb{E}[(Y-(aX+b))^2].$$

Usando técnicas de cálculo diferencial en varias variables el problema se reduce a resolver el sistema de ecuaciones $\nabla g = 0$. Es fácil ver, desarrollando cuadrados, que

$$\begin{aligned}\frac{\partial g(a,b)}{\partial a} &= 2a\mathbb{E}[X^2] - 2\mathbb{E}[XY] + 2b\mathbb{E}[X],\\ \frac{\partial g(a,b)}{\partial b} &= 2b - 2\mathbb{E}[Y] + 2a\mathbb{E}[X]\end{aligned}$$

El problema se reduce a resolver el siguiente sistema lineal de ecuaciones

$$\begin{cases} a\mathbb{E}[X^2] + b\mathbb{E}[X] = \mathbb{E}[XY] \\ a\mathbb{E}[X] + b = \mathbb{E}[Y] \end{cases}$$

Sumando la primera ecuación y la segunda multiplicada por $-\mathbb{E}[X]$, se obtiene

$$a(\mathbb{E}[X^2] - \mathbb{E}[X]^2) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \Leftrightarrow a = \frac{Cov(X,Y)}{\mathbb{V}(X)}.$$

Sustituyendo el valor de a en la segunda y despejando b se obtiene

$$b = \mathbb{E}[Y] - \frac{Cov(X,Y)}{\mathbb{V}(X)}\mathbb{E}[X].$$

Por lo tanto, *la recta de regresión de Y basada en X* es

$$\begin{aligned} \hat{Y} &= \frac{Cov(X,Y)}{\mathbb{V}(X)}X + \mathbb{E}[Y] - \frac{Cov(X,Y)}{\mathbb{V}(X)}\mathbb{E}[X] \\ &= \frac{Cov(X,Y)}{\mathbb{V}(X)}\left(X - \mathbb{E}[X]\right) + \mathbb{E}[Y]. \end{aligned} \tag{34}$$

Además el *error cuadrático medio* es igual a

$$\mathbb{E}[(Y - \hat{Y})^2] = \mathbb{V}(Y)\left(1 - \rho(X,Y)^2\right), \tag{35}$$

donde

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{36}$$

es el llamado *coeficiente de correlación* de las variables X, Y.

Coeficiente de correlación

El coeficiente de correlación definido en (36) es la covarianza de las variables normalizadas

$$X^* := \frac{X - \mathbb{E}[X]}{\sigma(X)}, \qquad Y^* := \frac{Y - \mathbb{E}[Y]}{\sigma(Y)}. \tag{37}$$

Este coeficiente es independiente de los orígenes y unidades de medida, esto es, para constantes a_1, a_2, b_1, b_2 con $a_1 > 0$, $a_2 > 0$, tenemos $\rho(a_1X + b_1, a_2Y + b_2) = \rho(X,Y)$.

Desafortunadamente, el término correlación sugiere implicaciones que no le son inherentes. Si X e Y son independientes, $\rho(X,Y) = 0$. Sin embargo la recíproca no es cierta. De hecho, *el coeficiente de correlación $\rho(X,Y)$ puede anularse incluso cuando Y es función de X.*

Ejemplo 3.2.

1. Sea X una variable aleatoria que toma valores $\pm 1, \pm 2$ cada uno con probabilidad $\frac{1}{4}$ y sea $Y = X^2$. La distribución conjunta está dada por

   $$p(-1,1) = p(1,1) = p(-2,4) = p(2,4) = 1/4.$$

   Por razones de simetría ($\mathbb{E}[X] = 0$ y $\mathbb{E}[XY] = 0$) $\rho(X,Y) = 0$ incluso cuando Y es una función de X.

2. Sean U y V variables *independientes* con la misma distribución, y sean $X = U + V$, $Y = U - V$. Entonces $\mathbb{E}[XY] = \mathbb{E}[U^2] - \mathbb{E}[V^2] = 0$ y $\mathbb{E}[Y] = 0$. En consecuencia, $Cov(X, Y) = 0$ y por lo tanto también $\rho(X, Y) = 0$. Por ejemplo, X e Y podrían ser la suma y la diferencia de los puntos de dos dados. Entonces X e Y son ambos pares ó ambos impares y por lo tanto dependientes.

Nota Bene. El coeficiente de correlación no es una medida general de la dependencia entre X e Y. Sin embargo, $\rho(X, Y)$ está conectado con la dependencia *lineal* de X e Y. En efecto, de la identidad (35) se deduce que $|\rho(X, Y)| \leq 1$ y que $\rho(X, Y) = \pm 1$ si y solo si Y es una función lineal de X (casí seguramente).

3.1. Ejemplo: Min y Max (dos dados)

El experimento consiste en arrojar dos dados equilibrados y observar las variables aleatorias definidas por $X =$"el mínimo de los dos resultados" e $Y =$"el mayor de los dos resultados".

El espacio de muestral asociado al experimento se puede representar en la forma $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$, donde cada punto $(i, j) \in \Omega$ indica que el resultado del primer dado es i y el resultado del segundo es j. Para reflejar que arrojamos *dos dados equilibrados*, todos los puntos de Ω serán equiprobables, i.e., para cada $(i, j) \in \Omega$ se tiene $\mathbb{P}(i, j) = 1/36$. Formalmente las variables aleatorias X e Y están definidas por

$$X(i, j) := \text{mín}\{i, j\} \qquad Y(i, j) := \text{máx}\{i, j\}. \tag{38}$$

Distribución conjunta y distribuciones marginales de X e Y. En primer lugar vamos a representar el espacio muestral Ω en la forma de una matriz para poder observar más claramente los resultados posibles

$$\Omega = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\} \tag{39}$$

Es claro que $p_{X,Y}(x, y) = 2/36$ para $x < y$; $p_{X,Y}(x, y) = 1/36$ para $x = y$ y que $p_{X,Y}(x, y) = 0$ para $y < x$. En el Cuadro 3 representamos la distribución conjunta $p_{X,Y}(x, y)$ y las distribuciones marginales p_X y p_Y. Como su nombre lo indica las marginales están en los márgenes de la tabla y se obtienen sumando filas y columnas.

Marginales. Observando el Cuadro 3 se deduce que las distribuciones marginales son

$$p_X(x) = \frac{13 - 2x}{36}, \qquad p_Y(y) = \frac{2y - 1}{36}.$$

$x \setminus y$	1	2	3	4	5	6	p_X
1	1/36	2/36	2/36	2/36	2/36	2/36	11/36
2	0	1/36	2/36	2/36	2/36	2/36	9/36
3	0	0	1/36	2/36	2/36	2/36	7/36
4	0	0	0	1/36	2/36	2/36	5/36
5	0	0	0	0	1/36	2/36	3/36
6	0	0	0	0	0	1/36	1/36
p_Y	1/36	3/36	5/36	7/36	9/36	11/36	

Cuadro 3: Distribución conjunta de (X, Y). En el margen derecho se encuentra la distribución marginal de X y en el margen inferior la marginal de Y.

(In)dependencia. Como la matriz de la distribución conjunta no es la tabla de multiplicar de las marginales las variables X e Y no son independiente. Lo que, por otra parte, constituye una obviedad.

Esperanzas.

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=1}^{6} x p_X(x) = \frac{1}{36}\sum_{x=1}^{6} x(13-2x) = \frac{91}{36}\\ \mathbb{E}[Y] &= \sum_{y=1}^{6} y p_Y(y) = \frac{1}{36}\sum_{y=1}^{6} y(2y-1) = \frac{161}{36},\end{aligned}$$

Nota Bene. Observar que $\mathbb{E}[Y] = 7 - \mathbb{E}[X]$. (*¿Por qué?*)

Covarianza. Recordamos primero que $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

$$\mathbb{E}[XY] = \sum_{x=1}^{6}\sum_{y=1}^{6} xy p_{X,Y}(x,y) = \frac{1}{36}\left(\sum_{x=1}^{6} x^2 + 2\sum_{1\le x<y\le 6} xy\right) = \frac{441}{36}.$$

En consecuencia,

$$Cov(X,Y) = \frac{441}{36} - \left(\frac{91}{36}\right)\left(\frac{161}{36}\right) = \frac{1225}{1296}.$$

Varianzas.

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{x=1}^{6} x^2 p_X(x) = \frac{1}{36}\sum_{x=1}^{6} x^2(13-2x) = \frac{301}{36},\\ \mathbb{V}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{301}{36} - \frac{8281}{1296} = \frac{2555}{1296}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Y^2] &= \sum_{y=1}^{6} y^2 p_Y(y) = \frac{1}{36}\sum_{y=1}^{6} y^2(2y-1) = \frac{791}{36},\\ \mathbb{V}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{791}{36} - \frac{25921}{1296} = \frac{2555}{1296}.\end{aligned}$$

Varianza de la suma.

$$\begin{aligned}\mathbb{V}(X+Y) &= \mathbb{V}(X) + \mathbb{V}(Y) + 2Cov(X,Y) = \frac{2555}{1296} + \frac{2555}{1296} + 2\left(\frac{1225}{1296}\right)\\ &= \frac{35}{6}.\end{aligned}$$

Nota Bene. Resultado por demás previsible. (*¿Por qué?*)

Condicionales. Dividiendo cada fila de la matriz $p_{X,Y}(x,y)$ por el correspondiente valor de su margen derecho se obtienen las funciones de probabilidad condicionales de Y dado que $X = x$.

$X \setminus Y$	1	2	3	4	5	6
1	1/11	2/11	2/11	2/11	2/11	2/11
2	0	1/9	2/9	2/9	2/9	2/9
3	0	0	1/7	2/7	2/7	2/7
4	0	0	0	1/5	2/5	2/5
5	0	0	0	0	1/3	2/3
6	0	0	0	0	0	1

Cuadro 4: Distribución condicional de Y dada X.

Los coeficientes de la fila x son la distribución condicional de Y dado que $X = x$. El coeficiente de la fila x y columna y es el valor de

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}.$$

Esperanza condicional. Para hallar $\mathbb{E}[Y|X]$ recordamos que la esperanza condicional es $\varphi(X)$, donde $\varphi(x) = \mathbb{E}[Y|X = x] = \sum_y y\, p_{Y|X}(y|x)$. Utilizando los datos de la fila x para calcular el valor de $\varphi(x)$, obtenemos lo siguiente

$$\varphi(1) = \frac{41}{11};\ \ \varphi(2) = \frac{38}{9};\ \ \varphi(3) = \frac{33}{7};\ \ \varphi(4) = \frac{26}{5};\ \ \varphi(5) = \frac{17}{3};\ \ \varphi(6) = \frac{6}{1}.$$

Se puede ver que

$$\varphi(x) = \frac{42 - x^2}{13 - 2x}\mathbf{1}\{1 \le x \le 6\}.$$

Por lo tanto,

$$\mathbb{E}[Y|X] = \frac{42 - X^2}{13 - 2X}.$$

Recta de regresión. La recta de regresión tiene la expresión

$$\hat{Y} = \frac{Cov(X,Y)}{\mathbb{V}(X)}(X - \mathbb{E}[X]) + \mathbb{E}[Y] = \frac{1225}{2555}\left(X - \frac{91}{36}\right) + \frac{161}{36} = \frac{245}{511}X + \frac{1666}{511}$$

Coeficiente de correlación. El coeficiente de correlación vale

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} = \frac{1225}{2555} \approx 0.4794...$$

□

4. Bibliografía consultada

Para redactar estas notas se consultaron los siguientes libros:

1. Billingsley, P.: Probability and measure. John Wiley & Sons, New York. (1986)
2. Bertsekas, D. P., Tsitsiklis, J. N.: Introduction to Probability. M.I.T. Lecture Notes. (2000)
3. Durrett R.:Probability.Theory and Examples. Duxbury Press, Belmont. (1996)
4. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons, New York. (1957)
5. Feller, W.: An introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York. (1971)
6. Maronna R.: Probabilidad y Estadística Elementales para Estudiantes de Ciencias. Editorial Exacta, La Plata. (1995)
7. Ross, S.: Introduction to Probability Models. Academic Press, San Diego. (2007)