Vectores aleatorios y covarianza

Kevin Piterman

26 de Mayo, 2020

Resumen.

Un vector aleatorio es una variable aleatoria del estilo $\mathrm{X} : \Omega \to \mathbb{R}^n$. Llamamos X_1, X_2, ..., X_n a sus coordenadas. Si $n = 2$ o 3 notamos X, Y, Z a las coordenadas del vector.

$$\mathrm{X} = (X, Y), \quad \text{o directamente} \quad (X, Y)$$

$$\mathrm{X} = (X, Y, Z), \quad \text{o directamente} \quad (X, Y, Z)$$

Función de distribución acumulada: $F_{\mathrm{X}}(A) = P(\mathrm{X} \in A) = P((X, Y) \in A)$, $A \subseteq \mathbb{R}^2$.

Discreta: si el vector toma un conjunto discreto de valores (finitos o numerables). Llamamos $R_{\mathrm{X}} = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots\}$ al rango de valores de X, y $p_{\mathrm{X}}(x_i, y_i) = P(\mathrm{X} = (x_i, y_i))$ es la función de probabilidad puntual.

Continuas: cuando existe una función de densidad $f_X : \mathbb{R}^n \to \mathbb{R}$ tal que $F_X(A) = \int_A f_{\mathrm{X}} dX$.

Marginales: cuando proyectamos cada coordenada de un vector, obtenemos las variables aleatorias X, Y, que tienen su propia distribución. Para conocer su distribución,

$$F_X(t) = P(X \leq t) = P(X \leq t, Y \in \mathbb{R}).$$

$\mathrm{X} = (X, Y)$	Discretas	Continuas
Función de	Probabilidad puntual p_{X}	Densidad f_{X}
Valores que toman	Rango R_{X} En general: $R_{\mathrm{X}} \subseteq R_X \times R_Y$	Soporte ($\mathrm{Supp}(f_{\mathrm{X}})$): donde mayormente no se anula f_{X}
Marginales	$p_X(x) = \sum_{(x,y) \in R_{\mathrm{X}}} p_{\mathrm{X}}(x, y)$ $= \sum_{y \in R_Y} p_{\mathrm{X}}(x, y)$	$f_X(x) = \int_{\mathbb{R}} f_{\mathrm{X}}(x, y) dy$
$\mathbb{E}(X, Y)$ (vector)	$\sum_{(x,y) \in R_{\mathrm{X}}} (x, y) p_{\mathrm{X}}(x, y)$	$\left(\int x f_{\mathrm{X}}(x, y) dx dy, \int y f_{\mathrm{X}}(x, y) dx dy\right)$
$h : \mathbb{R}^2 \to \mathbb{R}$, $\mathbb{E}(h(X, Y))$	$\sum_{(x,y) \in R_{\mathrm{X}}} h(x, y) p_{\mathrm{X}}(x, y)$	$\int h(x, y) f_{\mathrm{X}}(x, y) dx dy$
Distribuciones condicionales	$p_{X \mid Y=y}(x) = \frac{p_{\mathrm{X}}(x,y)}{p_Y(y)}$	$f_{X \mid Y=y}(x) = \frac{f_{\mathrm{X}}(x,y)}{f_Y(y)}$

X, Y independientes		
Función	(puntual) $p_{\mathrm{X}} = p_X . p_Y$	(densidad) $f_{\mathrm{X}} = f_X . f_Y$
Valores que toman: son rectángulos	$R_{\mathrm{X}} = R_X \times R_Y$	$\mathrm{Supp}(f_{\mathrm{X}}) = \mathrm{Supp}(f_X) \times \mathrm{Supp}(f_Y)$
$\mathbb{E}(X, Y)$ (vector)	$(\mathbb{E}(X), \mathbb{E}(Y))$	
Esperanza $\mathbb{E}(XY)$	$\mathbb{E}(X)\mathbb{E}(Y)$	
Distribuciones condicionales	$p_{X \mid Y=y} = p_X$	$f_{X \mid Y=y} = f_X$

Table 1: Cuadro resumen de vectores aleatorios.

Covarianza: $\text{Cov}(X,Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

- Idea: medir cuán correlacionadas están X, Y.
- En particular, si X, Y son independientes, $\text{Cov}(X,Y) = 0$, no están correlacionadas.
- **No vale la vuelta:** $\text{Cov}(X,Y) = 0$ no implica X, Y independientes.

Coeficiente de correlación: $\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$.

Idea geométrica de fondo:

- Cov es un producto interno del espacio vectorial de variables aleatorias "módulo constantes".
- No están correlacionadas (o sea que su covarianza sea 0) = "ortogonales".
- $\rho(X,Y) =$ "$\cos(\theta)$", el coseno del ángulo entre ambas.
- La varianza sería la norma al cuadrado: $V(X) = \text{Cov}(X,X)$.
- El desvío estándar sería la norma $\sigma = \sqrt{V(X)} = \sqrt{\text{Cov}(X,X)}$.

Propiedades. $a, b, c \in \mathbb{R}$ constantes.

1. (Constantes son 0) $\text{Cov}(X,a) = 0$.
2. (Simétrica) $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.
3. (Linealidad) $\text{Cov}(aX + bY, Z) = a\text{Cov}(X,Z) + b\text{Cov}(Y,Z)$.
4. (Definida positiva) $\text{Cov}(X,X) = V(X) \geq 0$, y es 0 si y solo si X es constante.

Interpretación. $\rho = \rho(X,Y)$

1. $\rho = 1$: entonces $Y = aX + b$ con $a > 0$. Correlación positiva perfecta. Las dos variables crecen proporcionalmente.
2. $0 < \rho < 1$: correlación positiva entre ellas, si una crece la otra también, pero no de manera proporicional.
3. $\rho = 0$: no están correlacionadas.
4. $-1 < \rho < 0$: correlación negativa entre ellas, si una crece la otra decrece, pero no de manera proporcional.
5. $\rho = -1$: entonces $Y = aX + b$ donde $a < 0$. Correlación negativa perfecta. Si una crece, la otra decrece proporcionalmente.

Ejemplo 1. Sean X, Y independientes tales que $V(X) = 4$ y $V(Y) = 1$. Calcular:

(a) $\text{Cov}(2X - 4Y, Y + X)$.

(b) $\rho(X, Z)$ si $Z = 2X - Y$.

(c) $V(XY)$ si $\mathbb{E}(X^2) = 1$ y $\mathbb{E}(Y^2) = 2$.

Solución.
(a) Aplicamos las propiedades de linealidad de la covarianza y resolvemos:

$$\begin{aligned}\text{Cov}(2X - 4Y, Y + X) &= 2\text{Cov}(X, Y) + 2\text{Cov}(X, X) - 4\text{Cov}(Y, Y) - 4\text{Cov}(Y, X) \\ &= 2V(X) - 4V(Y) && \text{son idptes} \\ &= 2.4 - 4.1 \\ &= 4.\end{aligned}$$

(b) Para calcular el $\rho(X, Z)$ necesitamos conocer $\text{Cov}(X, Z)$ y los desvíos $\sigma(X)$ y $\sigma(Z)$. Para calcular la covarianza y $\sigma(Z)$ utilizamos las propiedades de linealidad.

$$\text{Cov}(X, Z) = \text{Cov}(X, 2X - Y) = 2\text{Cov}(X, X) - \text{Cov}(X, Y) = 2V(X) = 8.$$

$$\sigma(Z)^2 = V(Z) = V(2X - Y) = 2^2V(X) + V(Y) - 2\text{Cov}(2X, Y) = 4.4 + 1 - 0 = 17.$$

Por lo tanto $\sigma(Z) = \sqrt{17}$. Como $\sigma(X) = 2$, tenemos que

$$\rho(X, Z) = \frac{\text{Cov}(X, Z)}{\sigma(X)\sigma(Z)} = \frac{8}{2\sqrt{17}} = \frac{4}{\sqrt{17}}.$$

(c) Utilizamos que X e Y son independientes y la definición de varianza/covarianza.

$$\begin{aligned}V(XY) = \text{Cov}(XY, XY) &= \mathbb{E}((XY)^2) - \mathbb{E}(XY)\mathbb{E}(XY) \\ &= \mathbb{E}(X^2Y^2) - \mathbb{E}(XY)\mathbb{E}(XY) \\ &= \mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(X)\mathbb{E}(Y)\mathbb{E}(X)\mathbb{E}(Y) && \text{son idptes} \\ &= 1.2 - (\mathbb{E}(X)^2\mathbb{E}(Y)^2) \\ &= 2 - (\mathbb{E}(X^2) - V(X))(\mathbb{E}(Y^2) - V(Y)) \\ &= 2 - (1 - 4)(2 - 1) \\ &= 2 - (-3) \\ &= 5.\end{aligned}$$

Ejemplo 2. Supongamos que queremos realizar un experimento que tiene r posibles resultados, cada uno con probabilidad $p_i \in (0,1)$ tal que $\sum_i p_i = 1$. Llevamos a cabo este experimento m veces de manera independiente. Sea N_i la cantidad de veces que salió el resultado i del experimento. Demostrar que N_i y N_j están negativamente correlacionadas e interpretar.

Solución. Cada N_i es una binomial de parámetros (m, p_i). Nosotros queremos calcular $\mathrm{Cov}(N_i, N_j)$ si $i \neq j$.

Como la suma de todas es m, o sea $N_1 + N_2 + \ldots + N_r = m$, vemos que si N_i aumenta, las demás tienen que disminuir su valor. Esto nos hace pensar que la correlación entre N_i y N_j, para $i \neq j$, es negativa.

Para calcular explícitamente su correlación, vamos a escribir a cada variable como suma de variables independientes ("ortogonales") de manera de simplificar el cálculo. Como estamos trabajando con variables binomiales, podemos escribirlas como sumas de variables de tipo Bernoulli, independientes.

Sean $I_k \sim Be(p_i)$ y $J_k \sim Be(p_j)$, para $1 \leq k \leq m$, de manera que I_k me dice si en el intento k-ésimo del experimento salió el resultado i o no (ídem con J_k).

Luego $N_i = \sum_k I_k$ y $N_j = \sum_l J_l$. Utilizando las propiedades de la covarianza, separamos la covarianza de N_i con N_j en suma de covarianzas entre variables de tipo Bernoulli.

$$\begin{aligned}
\mathrm{Cov}(N_i, N_j) &= \mathrm{Cov}\left(\sum_k I_k, \sum_l J_l\right) \\
&= \sum_{k,l} \mathrm{Cov}(I_k, J_l) \\
&= \sum_{k=1}^{m} \mathrm{Cov}(I_k, J_k) && \text{si } k \neq l \text{ son idptes} \\
&= \sum_{k=1}^{m} \mathbb{E}(I_k J_k) - \mathbb{E}(I_k)\mathbb{E}(J_k) && I_k J_k = 0 \\
&= \sum_{k=1}^{m} -\mathbb{E}(I_k)\mathbb{E}(J_k) \\
&= \sum_{k=1}^{m} -p_i p_j \\
&= -m p_i p_j.
\end{aligned}$$

Por lo tanto, como $\mathrm{Cov}(N_i, N_j) < 0$ están negativamente correlacionadas.

Interpretación: si una aumenta, la otra tiene que disminuir porque la cantidad de realizaciones del experimento se mantiene constante m.

Ejemplo 3. (Grafos aleatorios) Supongamos que tenemos n vértices numerados $1, 2, \ldots, n$. Sea $p \in (0, 1)$. Armamos el siguiente grafo aleatorio: ponemos una arista entre los vértices i y j (con $i \neq j$) con probabilidad p. Sea D_i la cantidad de aristas incidentes al vértice i. Notar que no estamos considerando aristas orientadas, las aristas son entre vértices diferentes, y no hay múltiples aristas entre dos vértices fijos.

(a) Determinar la distribución de D_i.

(b) Calcular $\rho(D_i, D_j)$ para todo i, j.

Solución.

(a) D_i es una binomial de parámetros $n-1$, p: porque puedo pensarlo como que me paro en el vértice i, y hago $n-1$ experimentos independientes donde cada uno de ellos corresponde a poner una arista entre el vértice i y el $j \neq i$ con probabilidad p. Como la cantidad de vértices distintos de i es $n-1$, este experimento se realiza $n-1$ veces y son todos independientes entre sí.

(b) Siguiendo las ideas del ejercicio anterior, queremos escribir como sumas de variables de tipo Bernoulli (independientes) a cada una de las variables binomiales D_i. Sean $E_{i,j} \sim Be(p)$ (si $i \neq j$) que indica si hay una arista entre i y j. Observemos que $E_{i,j} = E_{j,i}$. Ponemos $E_{i,i} = 0$ (porque no puede haber una arista del vértice i al i).

Entonces, fijados $i \neq j$,

$$D_i = \sum_k E_{i,k}, \qquad D_j = \sum_l E_{j,l}.$$

Calculamos su covarianza utilizando las propiedades de linealidad:

$$\mathrm{Cov}(D_i, D_j) = \sum_{k,l} \mathrm{Cov}(E_{i,k}, E_{j,l}).$$

Como dos aristas que salen de i y de j respectivamente son distintas si caen en vértices distintos, ponerlas o no ponerlas es independiente entre ellas, y por lo tanto $\mathrm{Cov}(E_{i,k}, E_{j,l}) = 0$ si $(i,k) \neq (j,l)$ o (l,j). Entonces el único caso que tenemos que considerar es $k = j$ y $l = i$. Por lo tanto,

$$\mathrm{Cov}(D_i, D_j) = \mathrm{Cov}(E_{i,j}, E_{j,i}) = V(E_{j,i}) = V(Be(p)) = p(1-p).$$

Además, $\sigma(D_i) = \sigma(D_j) = \sqrt{(n-1)p(1-p)}$, y así

$$\rho(D_i, D_j) = \frac{p(1-p)}{(n-1)p(1-p)} = \frac{1}{n-1}.$$

Finalmente, en el caso $i = j$, $D_i = D_j$ y la correlación es perfecta positiva, o sea $\rho(D_i, D_j) = 1$.